Methods for Personalized and Evidence Based Medicine

Zachary Samuelson Shahn

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

Columbia University
2016

Methods for Personalized and Evidence Based Medicine
Zachary Samuelson Shahn

Abstract

There is broad agreement that medicine ought to be 'evidence based' and 'personalized' and that data should play a large role in achieving both these goals. But the path from data to improved medical decision making is not clear. This thesis presents three methods that hopefully help in small ways to clear the path.

Personalized medicine depends almost entirely on understanding variation in treatment effect. Chapter 1 describes latent class mixture models for treatment effect heterogeneity that distinguish between continuous and discrete heterogeneity, use hierarchical shrinkage priors to mitigate overfitting and multiple comparisons concerns, and employ flexible error distributions to improve robustness. We apply different versions of these models to reanalyze a clinical trial comparing HIV treatments and a natural experiment on the effect of Medicaid on emergency department utilization.

Medical decisions often depend on observational studies performed on large longitudinal health insurance claims databases. These studies usually claim to identify a causal effect, but empirical evaluations have demonstrated that standard methods for causal discovery perform poorly in this context, most likely in large part due to the presence of unobserved confounding. Chapter 2 proposes an algorithm called Ensembles of Granger Graphs (EGG) that does not rely on the assumption

that unobserved confounding is absent. In a simulation and experiments on a real claims database, EGG is robust to confounding, has high positive predictive value, and has high power to detect strong causal effects.

While decision making inherently involves causal inference, purely predictive models aid many medical decisions in practice. Predictions from health histories are challenging because the space of possible predictors is so vast. Not only are there thousands of health events to consider, but also their temporal interactions. In Chapter 3, we adapt a method originally developed for speech recognition that greedily constructs informative labeled graphs representing temporal relations between multiple health events at the nodes of randomized decision trees. We use this method to predict strokes in patients with atrial fibrillation using data from a Medicaid claims database.

I hope the ideas illustrated in these three projects inspire work that someday genuinely improves healthcare. I also include a short 'bonus' chapter (based mostly on the prior work of Li Ye, a former student of my advisor) on an improved estimate of effective sample size in importance sampling. This chapter is not directly related to medicine, but finds a home in this thesis nonetheless.

Contents

List of Figures

List of Tables

Chapter 1

Latent Class Mixture Models of Treatment Effect Heterogeneity

1.1 Introduction

In randomized experiments, it is often of interest to characterize treatment effect heterogeneity in terms of baseline covariates. Usually, the aim is to identify subpopulations likely to have particularly positive or negative (or neutral) responses to treatment. The process of searching for such subpopulations after the completion of an experiment (without pre-specifying which subpopulations will be considered as candidates) is called 'post hoc subgroup analysis'. It is a controversial practice. Concerns about data dredging and multiple comparisons (Rothwell, 2005) have led many authors to advise against reporting results from post hoc subgroup analyses at all. However, it is our view that post hoc analyses can produce informative insights that would be unlikely to arise from limited pre-registered comparisons. Here we illustrate an approach in which identification of special subgroups is one byproduct of fully modeling treatment effect heterogeneity more generally. By placing shrinkage priors on relevant parameters and applying cross validation based

tools for model evaluation and comparison, we minimize data dredging concerns and provide a mechanism for gauging confidence in the substantive implications of model results.

Employing parametric probability models of heterogeneity brings certain automatic advantages. The parameter estimates have interpretable implications about the shape of heterogeneity, and models provide estimates of uncertainty about those parameters. Models also provide estimates of treatment effects for subpopulations and corresponding uncertainty estimates. And Bayesian models in particular allow us to place hierarchical shrinkage priors on relevant parameters to avoid overfitting/data dredging (Gelman et al., 2012). These are obvious features of parametric probability models and are only worth mentioning because many methods for subgroup analysis are nonparametric or not model based and do not share these features.

Of course, any model is bound to be misspecified, and the advantages described above only apply insofar as the chosen model is a good approximation to reality. We choose a relatively flexible class of models to improve the chances that some model in the class approximates reality well. We employ Leave One Out Cross Validation (LOO-CV) techniques for model comparison and evaluation (Gelfand, 1996; Vehtari and Lumpenin, 2001). For both prediction and inference we prefer models with better LOO-CV estimated expected utility, where a good choice for utility in this context is posterior predictive density. We estimate uncertainty about the expected utility of each candidate model and adjust our degree of belief in each model's implications accordingly. If multiple models could plausibly have

the best true expected utility, we do not draw firm conclusions about aspects of heterogeneity on which those models disagree. If one model is clearly superior to the rest and appears to fit adequately based on posterior predictive checks (Gelman et al., 1996), then we would be fairly confident in its implications.

Specifically, we propose to model treatment effect heterogeneity using regularized Bayesian latent class mixture models with treatment interaction terms and flexible error distributions. Such models are particularly well suited to illuminate the shape of heterogeneity. The treatment interaction terms capture 'continuous heterogeneity' while the latent class components capture 'discrete heterogeneity.' By continuous heterogeneity we mean variation in subjects' individual treatment effects that is well approximated by a smooth function of underlying covariates. Discrete heterogeneity refers to variation in subjects' individual treatment effects that is associated with latent class membership, where latent class membership may in turn be associated with baseline covariates. Discrete heterogeneity is likely to be present if a treatment works through unobserved causal pathways that may be discretely open or closed. For example, suppose a drug works better in people with a specific phenotype for some protein receptor, but the presence of that phenotype is not recorded as a baseline covariate in a clinical trial evaluating the drug. Further, suppose that a recorded baseline covariate (say, weight) is associated with the presence of the beneficial phenotype. Then treatment effect variation as a function of weight will be better approximated by a latent class model with weight as a predictor of latent class membership than by any smooth function of weight alone. It can sometimes be important to understand which type of heterogeneity is present.

Despite our approach being a fairly straightforward application of Bayesian latent class mixture models and existing model comparison and evaluation techniques, we have not seen it in the subgroup analysis literature. Further, our approach offers a different combination of strengths (and weaknesses) from those methods we have seen.

One general tactic in the literature is to use nonparametric machine learning algorithms (often based on trees) to predict counterfactual outcomes of future subjects under treatment and control. Examples of works in this vein include (Kang et al, 2012; Foster et al., 2010; Su et al, 2009; and others). These methods likely have the edge over ours when it comes to predictive accuracy and flexibility. Some of them also use cross validation to effectively mitigate multiple comparisons concerns. However, many do not provide estimates of uncertainty about their predictions and usually do not characterize the shape of heterogeneity interpretably. (Athey and Imbens, 2015) recently proposed a machine learning approach that produces valid standard errors for causal effect estimates within nodes of a tree fit to a holdout validation set. Of course, holdout validation sets may not be practical for smaller experiments.

A closely related line of work directly learns optimal treatment assignment rules without first estimating counterfactual outcome response surfaces (Qian and Murphy, 2011; Zhang et al, 2012; Zhao et al, 2012). These methods are not interested in learning about heterogeneity, just assigning the best treatment to each subject. They have similar strengths and weaknesses relative to our method as the machine

learning approaches described above.

(Imai and Ratkovic, 2012) employ a linear SVM model with interaction terms to model heterogeneity. The output of their model is interpretable, and they place shrinkage penalties on the parameters to discourage overfitting. They use a cross validation measure for model selection. One could replace their SVM with a regression probability model and obtain uncertainty estimates as well. However, they do not directly model discrete heterogeneity and do not consider uncertainty in their model selection criterion.

There have been other examples of latent class mixture models in the literature. In another context, (Sobel and Muthen, 2012) used a logistic-normal latent class mixture model to reflect the assumption that there exists a subpopulation in which the treatment has zero effect. (Shen and He, 2015) recently applied a similar model to identify subgroups and developed a corresponding likelihood ratio test for the existence of latent treatment effect classes. Neither of these approaches allows for continuous effect modification, however, and both are very sensitive to the assumption of a normal error distribution. We use very flexible error distributions so that our estimates are robust to departures from normality. Working in a model evaluation and comparison framework as opposed to Shen's and He's hypothesis testing framework allows us to consider more complex models leading to better fits and more reliable results.

The structure of this chapter is as follows. In section 2, we describe our approach in detail, providing specifications of various models that we consider and explain-

ing the cross validation approach to model selection. In section 3, we provide simulations illustrating the utility of our approach and the importance of some of its features. In section 4, we reanalyze a clinical trial for an HIV treatment that was used as an example in (Shen and He, 2015). We conclude that this trial exhibits strong discrete heterogeneity, but not as strong as estimated by (Shen and He, 2015). In section 5, we apply our approach to data from the Oregon Health Insurance Experiment (OHIE). The OHIE was a natural experiment that arose when Oregon instituted a lottery to determine who could enroll in a new Medicaid program with limited openings. This experiment allowed researchers to explore various public health and economic effects of Medicaid. One prominent finding was that Medicaid increased emergency department (ED) utilization contrary to many experts' predictions. The researchers performed multiple pre-registered and post hoc comparisons between subgroups and discovered several possible heterogeneities. Because the OHIE study contains lots of noncompliers (i.e. lottery winners who did not enroll in Medicaid and lottery losers who managed to enroll through other channels), we follow the researchers in performing an instrumental variable analysis using the principal stratification framework of (Imbens and Rubin, 1996). Our method extends naturally to this framework because principal strata are themselves latent classes. When we include all covariates and employ hierarchical shrinkage priors to deal with multiple comparisons, we do not see strong evidence of heterogeneity associated with any of the observed covariates. In section 5, we conclude.

Before proceeding, we prominently note a major limitation. Latent class mixture models are not identifiable for categorical outcome distributions such as the

Bernoulli (Titterington, 1985). They are identifiable for the Poisson, negative binomial, or almost any continuous outcome distribution, however (Titterington, 1985).

1.2 Methods

1.2.1 Potential Outcomes

We use the potential outcomes framework (Rubin, 1974; Neyman, 1923) which formalizes the notion that each experimental unit (e.g. patient) has a potential outcome for each possible treatment assignment that unit might have received. We get to observe only the potential outcome corresponding to the treatment actually received. We consider the counterfactual potential outcome that would have been observed had the treatment assignment been different to be an unobserved random variable. For the i^{th} unit, let $Z_i \in \{0, 1\}$ denote the treatment assignment and $Y_{z,i}$ the potential outcome corresponding to the possibly counterfactual treatment assignment $Z_i = z$. Then $Y_{z=1,i} - Y_{z=0,i}$ is the effect of treatment on unit i. Note that this individual level causal effect can never be observed because we never observe both potential outcomes. Still, if we specify a model for the observed data it is possible to estimate $E[Y_{z=1,i} - Y_{z=0,i}|X_i]$ – the conditional average treatment effect (or CATE) – and the parameters that govern its value as a function of covariate values X_i.

1.2.2 Model Specification

In Figure 1, we provide a graphical specification that describes all the models we consider in this paper.

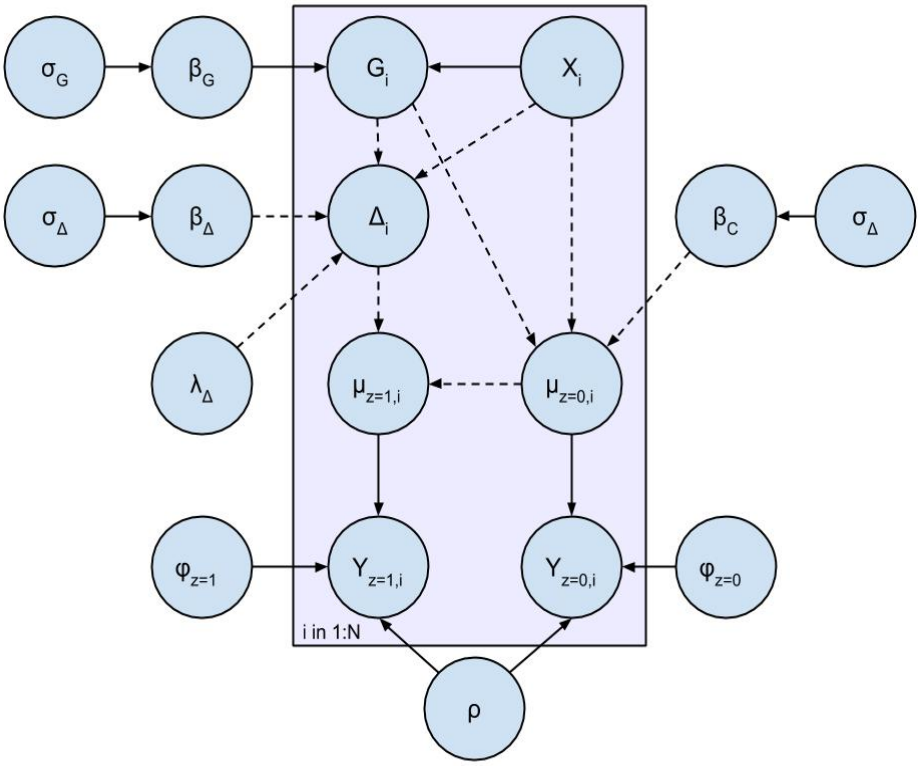

Figure 1.1: Graphical model specification. The dashed lines indicate deterministic relationships, and the solid lines indicate stochastic relationships.

Intent to Treat Analyses

An interpretation of the nodes of the graph in Figure 1 in the context of a standard Intent To Treat (ITT) analysis is as follows. The potential outcomes for the i^{th} unit under treatment assignment $Z = z$ for z in $\{0, 1\}$ are denoted $Y_{z,i}$. Again, for each patient we only observe one of these and the other is treated as missing.

The expected potential outcome for a unit with covariates X_i and latent class G_i under treatment assignment $Z = z$ is denoted by $\mu_{z,i}$. That is, the $\mu_{z,i}$ are the marginal expectations of the potential outcomes conditional on covariates and latent class. ϕ_z are parameters governing the marginal distributions of the potential outcomes $Y_{z,i}$ apart from their means $\mu_{z,i}$. For example, in a normal model, the ϕ_z would be standard deviations. ρ governs the dependence between the two potential outcomes but is completely unidentified because we never observe both potential outcomes for any one unit. $\mu_{z=0}$ is a function of covariates X_i, latent class G_i, and parameters β_C. $\mu_{z=1,i} = \mu_{z=0,i} + \Delta_i$, where Δ_i denotes the average treatment effect for units with covariates X_i and latent class G_i. Δ_i is a function of X_i, G_i, and parameters β_Δ and λ_Δ. β_Δ determines how treatment effect varies continuously as a function of covariates, and λ_Δ determines the magnitudes of the discrete differences in treatment effect between latent classes. The probability distribution that generates G_i is a function of X_i and parameters β_G. β_Δ, λ_Δ, and β_G are the parameters of interest as together they describe how treatment effect heterogeneity is related to covariates. The parameters σ_G, σ_Δ, and σ_C are variances for shrinkage priors that we place on relevant parameters to avoid overfitting. We put weakly informative priors on the shrinkage variances themselves so that the appropriate level of shrinkage is learned from the data.

We make a few remarks on identifiability. We have included the dependence parameter ρ in the graphical model even though it is completely unidentified by data. This is because we do not wish to assert that the potential outcomes are conditionally independent given X and G, as would be implied if ρ were not in the graph. An important consequence of the unidentifiability of ρ is that it is im-

possible to obtain an informative posterior or posterior predictive distribution for an individual level causal effect without making unverifiable assumptions about the dependence between potential outcomes. We are limited to inferences and predictions involving only parameters governing the marginal distributions of the potential outcomes. These parameters are identified by the data, and their posterior distributions are not impacted by ρ (Chib, 2007). Suppose, for example, we want to predict the causal effect of a drug on a new patient using a heterogeneity model of the sort sketched above fit to the clinical trial for the drug. The most we can extract from from this (or any) model without making assumptions about ρ is a posterior predictive distribution of the average treatment effect for patients with the same covariates and (redundantly, unobserved) latent class as our new patient (i.e. the posterior predictive distribution of that new patient's Δ parameter). We can also obtain marginal posterior predictive distributions for each of that patient's potential outcomes but not their difference (i.e. the patient's treatment effect). Finally, to avoid aliasing issues in parameters of interest, in all models of this form we require that $\lambda_{\Delta}^{G_i}$ increases with the latent class label G_i.

The framework of the graphical model in Figure 1 allows for flexibility in the selection of functional forms, distributions, and number of latent classes. In this chapter, we only consider linear models for the potential outcomes and logistic regression models for latent class membership. For continuous outcomes, we consider models with two different error distributions–M_{Norm} with a normal error distribution and M_{Flex} with a more flexible three component Gaussian mixture error distribution. We demonstrate through simulations in section 3 that violations of normality in M_{Norm} can lead to biased estimation of parameters of interest, but

using flexible error distributions as in M_{flex} solves this problem. M_{Norm} and M_{Flex} are specified below for the case of two latent classes, including application specific weakly informative priors. Extension to multinomial logistic regression latent class models is straightforward.

$\mathbf{M}_{Norm}$ specification:

$$
\begin{aligned}
&Y_{z=0,i} \sim N(\mu_{z=0,i}, \sigma_{z=0}) \\
&Y_{z=1,i} \sim N(\mu_{z=1,i}, \sigma_{z=1}) \\
&G_i \sim Bernoulli(p_i) \\
&logit(p_i) = \alpha_G + X_i\beta_G \\
&\mu^i_{z=0} = \alpha_C^{G_i} + X_i\beta_C^{G_i} \\
&\Delta_i = \lambda_\Delta^{G_i} + X_i\beta_\Delta \\
&\mu^i_{z=1} = \mu^i_{z=0} + \Delta_i \\
&\beta_C \sim N(0, \sigma_C) \\
&\beta_G \sim N(0, \sigma_{\beta_G}) \\
&\beta_\Delta \sim N(0, \sigma_\Delta) \\
&\sigma_C \sim Uniform(0, 10) \\
&\sigma_G \sim Uniform(0, 5) \\
&\sigma_\Delta \sim U(0, 10) \\
&\sigma_1 \sim U(0, 10) \\
&\sigma_0 \sim U(0, 10) \\
&\lambda^1_\Delta \sim N(0, 1000) \\
&\lambda^2_\Delta - \lambda^1_\Delta \sim Truncated\_Normal(0, 1000; 0+)
\end{aligned}
\tag{1.1}
$$

$\mathbf{M}_{Flex}$ is the same as M_{Norm} except the M_{Flex} error distributions are each a

mean 0 mixture of three Gaussian components. In $\mathbf{M}_{Flex}$,

$$
\begin{aligned}
&Y_{z,i} \sim q_1^z N(\mu_{z,i} + d_1^z, \sigma_1^z) + q_2^z N(\mu_{z,i} + d_2^z, \sigma_2^z) + q_3^z N(\mu_{z,i} + d_3^z, \sigma_3^z) \\
&\sum q_i^z = 1 \\
&\sum q_i^z d_i^z = 0 \\
&d_1 \sim N(0, 1000) \\
&d_2 \sim N(0, 1000) \\
&d_3 \text{ is determined by the constraint that the error distribution has mean 0} \\
&\mathbf{q} \sim Dirichlet(1)
\end{aligned}
\tag{1.2}
$$

Aliasing can arise in the estimation of the parameters governing the flexible error distribution, but that is not a problem because we are not interested in interpreting those parameters.

Instrumental Variable Analyses

Sometimes a situation arises in which treatment is not randomly assigned, but an encouragement to take treatment is randomly assigned. If the random encouragement only affects the outcome through the treatment and is indeed effective at inducing some people to take the treatment, then the encouragement is referred to as an 'instrument' and an Instrumental Variable (IV) analysis may be performed. IV analyses estimate the treatment effect in the subpopulation of units that would take the treatment if and only if encouraged by their value of the instrument. Such units are referred to as 'compliers' and the causal estimand in an IV analysis is referred to as the Complier Average Causal Effect (CACE). The canonical example

of an IV setting is a randomized clinical trial with noncompliance. It is frequently the case in clinical trials that participants do not comply with their treatment assignments. Patients in the treatment arm may fail to take the treatment, and those in the control arm may find a way to take the treatment anyway. Thus, a simple ITT analysis comparing the two arms of the trial estimates the effect of treatment assignment rather than the effect of the treatment itself. However, random assignment to the treatment arm can be viewed as an instrument that encourages patients to take the treatment. An IV analysis with assignment as the instrument then estimates the effect of treatment on those patients who would comply with whatever random treatment assignment they happened to receive.

We follow (Sobel and Muthen, 2012) in extending latent class heterogeneity to an instrumental variable (IV) setting. We consider the case of a randomly assigned binary instrument Z that encourages a binary treatment D. Suppose without loss of generality that $Z = 1$ encourages $D = 1$. The outcome is denoted by Y. We use a potential outcomes framework modified for the IV setting. As before, each subject is assumed to have a potential outcome for each possible treatment assignment (i.e. $Y_{D=1}$ and $Y_{D=0}$). Each subject is also assumed to have a potential treatment assignment for each possible value of the instrument (i.e. $D_{Z=1}$ and $D_{Z=0}$). Further, each subject has a potential outcome for each possible instrument value (i.e. $Y_{Z=1}$ and $Y_{Z=0}$). We assume that Z only affects the outcome Y through D, so

$$Y_{Z=z} = Y_{D_{Z=z}}. \tag{1.3}$$

Under this assumption, the CACE is equivalent to the average causal effect of the instrument Z on Y among compliers. That is,

$$CACE = E[Y_{D=1} - Y_{D=0} | D_{Z=1} = 1, D_{Z=0} = 0] = E[Y_{Z=1} - Y_{Z=0} | D_{Z=1} = 1, D_{Z=0} = 0] \tag{1.4}$$

In other words, estimating the CACE amounts to estimating the causal effect of the instrument among a subgroup (compliers). We are therefore interested in modeling the heterogeneity of the effect of the instrument within the (latent) subgroup of compliers. This places us back in a similar position to the ITT case.

Indeed, models of treatment effect heterogeneity in the instrumental variable setting can be represented by the same graphical model (Figure 1) as the standard ITT case. However, the interpretation of certain nodes changes, and there are certain added constraints on parameter values. Latent class (the G_i node in Figure 1) now encodes compliance status as well as treatment effect class. We follow (Imbens and Rubin, 1996) in defining four types of subjects or 'principal strata': always takers, never takers, compliers, and defiers. Their definitions are as follows: always takers would take the treatment regardless of their instrument value; never takers would not take the treatment regardless of their instrument value; compliers would take the treatment if and only if encouraged by their instrument; and defiers would take the treatment if and only if discouraged by their instrument. We make the common assumption that there are no defiers. We get to observe the principal strata of some subjects, but other subjects' principal strata are latent. Units with $Z = 1$ and $D = 0$ are definitely never takers, and units with $Z = 0$ and $D = 1$ are definitely always takers. But units with $Z = 1$ and $D = 1$ could

either be compliers or always takers, and units with $Z = 0$ and $D = 0$ could either be compliers or never takers. G_i takes one value for never takers, one value for always takers, and one value for each treatment effect class for compliers to allow for discrete heterogeneity in the CACE. Because we assume that the instrument only affects the outcome through the treatment, the instrument effect (represented by Δ_i in Figure 1) must be 0 whenever latent class G_i indicates a never-taker or always-taker.

In the IV application we consider in this paper, the outcome (number of visits to the emergency department) is a count variable which we modeled as negative binomial. We parameterized the outcome in terms of its mean and allowed the log of the mean to vary discretely with latent class and continuously with covariates. We call the resulting model M_{IV}, and it is specified below (including application specific weakly informative priors).

$\mathbf{M}_{IV}$ specification:

$$
\begin{aligned}
&Y^i_{z=0} \sim NegBinom(p^i_{z=0}, r_{z=0}) \\
&Y^i_{z=1} \sim NegBinom(p^i_{z=1}, r_{z=1}) \\
&G_i \sim \text{Multinomial\_Logistic\_Regression}(X_i; \beta_G) \\
&p^i_{z=0} = r_{G_i}/(r_{G_i} + \mu^i_{z=0}) \\
&p^i_{z=1} = r_{G_i}/(r_{G_i} + \mu^i_{z=1}) \\
&log(\mu^i_{z=0}) = \alpha^{G_i}_C + X_i\beta^{G_i}_C \\
&\Delta_i = \lambda^{G_i}_\Delta + X_i\beta^{G_i}_\Delta, \text{ where } \lambda^{G_i}_\Delta \text{ and } \beta^{G_i}_\Delta \text{ are set to 0} \\
&\qquad\qquad\qquad\qquad \text{when } G_i \text{ is Never Taker or Always Taker} \\
&log(\mu^i_{z=1}) = log(\mu^i_{z=0}) + \Delta_i \\
&\beta_C \sim N(0, \sigma_C) \\
&\beta_G \sim N(0, \sigma_{\beta_G}) \\
&\beta_\Delta \sim N(0, \sigma_\Delta) \\
&\sigma_C \sim Uniform(0, 10) \\
&\sigma_G \sim Uniform(0, 5) \\
&\sigma_\Delta \sim U(0, 10) \\
&r_1, \ldots, r_M \sim U(0, 10) \text{ where } M \text{ denotes the number of latent classes} \\
&\lambda^{complier,1}_\Delta \sim N(0, 5) \\
&\lambda^{complier,2}_\Delta - \lambda^{complier,1}_\Delta \sim Unif(0, 5)
\end{aligned}
\tag{1.5}
$$

Identifiability Issues

Identifiability issues can arise for any of the models discussed above when in reality there are no latent classes. In this case, if the error distributions are properly specified, there can be negligible or no difference in the likelihood between different parameter settings. For instance, the value of α_G (which determines probability of class membership) is irrelevant if there is no difference between classes. The estimates of β_G will still converge to 0 if there are no latent classes, though, so there is not danger of wrongly concluding that there is discrete heterogeneity *associated with observed covariates*.

If the error distributions are misspecified, the latent class component of the model might help to better model them. If there are no latent classes in reality, an MCMC may still converge to unique parameter values that best model the misspecified error distributions. Despite convergence, it is still not correct to interpret latent class component parameters in terms of heterogeneity in this scenario. But, again, if there are no latent classes in reality then the estimates of β_G should be near 0 and there is not danger of wrongly concluding that heterogeneity is associated with observed covariates.

Even if there are latent classes in reality, improvements in likelihood from better modeling misspecified error distributions can pull estimates away from their 'correct' values (that is, the values with correct implications about heterogeneity if interpreted as intended). That is why it is important to include flexible error distributions in the model. Simulations in Section 3.2 illustrate this phenomenon.

Model Evaluation

We follow the framework for model comparison by Bayesian cross validation laid out in (Vehtari and Lampinen, 2002). A sensible measure of a model M's value is the expected utility of using M to make predictions about future observations generated by the same process that generated the training data. A Bayesian model produces a posterior predictive distribution for future outcome y_{new} given future covariates x_{new} and the data D that the model M was fit to:

$$p(y|x_{new}, D, M) = \int p(y|x_{new}, \theta, D, M)p(\theta|D, M)d\theta, \tag{1.6}$$

where θ denotes the model parameters. The utility of M for predicting a new outcome is some function $u[y_{new}, x_{new}, p(y|x_{new}, D, M)]$ of the outcome and the posterior predictive distribution that measures how well the posterior predictive distribution predicted the outcome. We can estimate the expected utility of a model on populations similar to the training data as the average

$$\frac{1}{N}\sum_{i=1}^{N} u[y_i, x_i, p(y_i|x_i, D^{-i}, M)] \tag{1.7}$$

where D^{-i} denotes the data with the i^{th} observation removed and N is the number of observations in D. This is the LOO-CV estimate of the expected utility of a model M. To estimate the expected utility on a population whose covariates differ from the training data in known ways, a weighted average can be used. Because it is computationally prohibitive to fit the model once for each data point, we approximate the LOO-CV estimate using an importance sampling scheme proposed by (Gelfand, 1996; Vehtari and Lampinen, 2002). To compare two models M_1 and

M_2, interest lies in their expected difference in utility, which can be estimated as:

$$\bar{u}_{M_1-M_2} = \frac{1}{N}\sum_{i=1}^{N} u[y_i, x_i, p(y_i|x_i, D^{-i}, M_1) - u[y_i, x_i, p(y_i|x_i, D^{-i}, M_2)]. \quad (1.8)$$

Generally, the choice of utility function depends on the application. For many applications, the posterior predictive mean is taken as the forecast and an appropriate utility is a monotonic function of the distance of the posterior predictive mean from the actual outcome. For example, the squared error utility function would be:

$$u_{se}[y_{new}, x_{new}, p(y|x_{new}, D, M)] = (y_{new} - \int yp(y|x_{new}, D, M)dy)^2. \quad (1.9)$$

Such utilities are problematic for the purpose of distinguishing models that contain discrete latent class heterogeneity from those that contain only continuous heterogeneity because they ignore the shape of the posterior predictive distribution. If there really is heterogeneity, the posterior predictive distribution for a latent class model will be multimodal and its mean will lie somewhere between the modes. The posterior predictive distribution for a continuous effect modification model will usually be unimodal but have a similar mean, so utilities based on the accuracy of the posterior predictive mean will have low power to distinguish these potentially quite different models.

A commonly used utility function that does not suffer from this problem is the

posterior predictive density (ppd):

$$u_{ppd}[y_{new}, x_{new}, p(y|x_{new}, D, M)] = p(y_{new}|x_{new}, D, M). \tag{1.10}$$

This utility rewards models that place lots of posterior predictive probability mass near future outcome values. A model with a multimodal posterior predictive distribution would be rewarded for outcomes that lie near any mode and penalized for outcomes that lie in the low density regions between modes. This utility has several nice theoretical properties as well. The model with the highest mean posterior predictive density minimizes Kullback Leibler distance to the true model. Posterior predictive density is also a proper scoring rule (Dawid and Musio, 2014). A drawback of this utility is that it is sensitive to our choice of error distribution, and we do not directly care about modeling the error distribution for our application. If we use very flexible error distributions for all candidate models, though, this should not be a serious problem.

In practice, for any given experiment we might consider many candidate models $M_1, \ldots, M_K$ with varying numbers of latent classes and functional forms. We prefer the model with the highest LOO-CV estimated expected posterior predictive density. However, we want to be mindful of the possibility that, due to sampling variability, the model with the highest estimated expected utility is not the model with the highest true expected utility. Each model's estimated expected utility is the sample mean of the LOO-CV posterior predictive densities of all the observations from the experiment. Since the samples of LOO-CV ppds produced by each model are based on the same observations, they are dependent and their centers

can be compared using classical methods for dependent samples such as paired t-tests or Wilcoxon signed rank tests. Suppose that M_i has the highest estimated expected utility. We can obtain a conservative p-value for the null hypothesis that M_i has the highest true expected utility of all models considered by taking the p-value of the comparison between M_i and the next best model and adjusting for K multiple comparisons using Holm's method (Holm, 1979). There are K possible comparisons we might have made because we would have tested this null hypothesis for whichever model had the best estimated utility. If the p-value we obtain in this way is very low, we would weight the implications of the top model highly compared to the other candidates. If the p-value is high, we would not dismiss the implications of other models with comparable utilities and would accept uncertainty where those implications conflicted with our chosen model.

Of course, just because a model is the best of those we considered does not mean it is a good model. We perform posterior predictive checks (Gelman and Xiao, 1995; Bayesian Data Analysis) to try to identify deviations of our chosen model from the data. If we fail to identify any serious lack of fit, this improves confidence in the conclusions we draw from our model. If we do identify lack of fit, we can address them with new models and repeat the process described above.

1.3 Simulations

We apply our approach in several simulated examples demonstrating its capabilities and the importance of some of its features. First, we demonstrate the ability of cross validation to distinguish between discrete and continuous heterogeneity.

Next, we illustrate the necessity of flexible error distributions. All code for simulations discussed in this section is available at zshahn.columbia.edu.

In Sim_1, we generated data from M_{Norm} with the following settings:

$$
\begin{aligned}
&Y_{z=0,i} \sim N(\mu_{z=0,i}, 1) \\
&Y_{z=1,i} \sim N(\mu_{z=1,i}, 1) \\
&G_i \sim Bernoulli(p_i) \\
&logit(p_i) = -1 + X_i \cdot (-2, -1, 0, 1, 2) \\
&\mu^i_{z=0} = 5\mathbb{1}\{G_i = 2\} + X_i \cdot (-2, -1, 0, 1, 2) \\
&\Delta_i = 5\mathbb{1}\{G_i = 1\} + 15\mathbb{1}\{G_i = 2\} + X_i \cdot (-2, -1, 0, 1, 2) \\
&\mu^i_{z=1} = \mu^i_{z=0} + \Delta_i
\end{aligned}
$$

We simulated a two armed clinical trial with 500 patients in each arm. X consisted of 5 predictor variables generated from a standard normal distribution. Both continuous and discrete heterogeneity was present. We then fit three models to this data: M_{Flex}, $M_{Flex}^{Continuous}$ (which is identical to M_{Flex} but without a discrete heterogeneity component), and $M_{Flex}^{Constant}$ (which is identical to M_{Flex} but without discrete or continuous heterogeneity). Below, we see that the posterior of M_{Flex} is accurate and clearly does not sacrifice too much precision for the robustness gained from flexible error distributions.

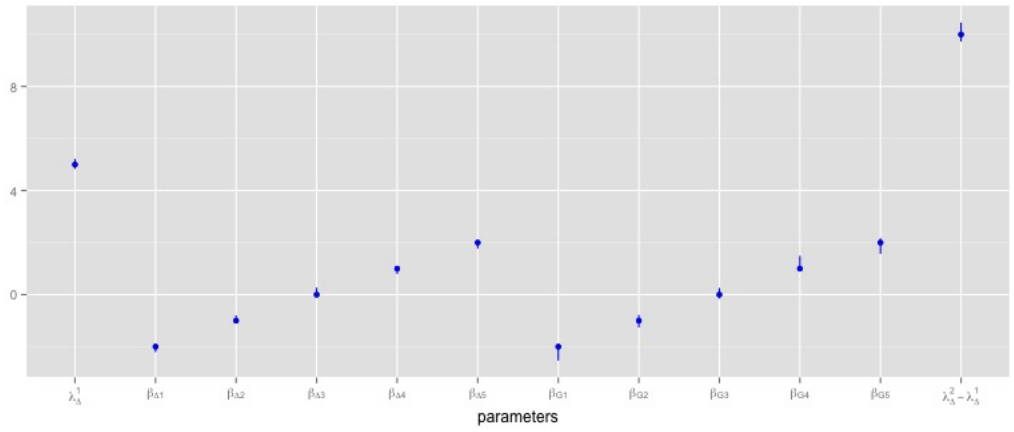

Figure 1.2: The results of fitting M_{Flex} to Sim_1. The dots are the true parameter values and the lines are 95% credible intervals.

The LOO-CV estimated expected posterior predictive densities for M_{Flex}, $M_{Flex}^{Continuous}$, and $M_{Flex}^{Constant}$ were 0.22, 0.12, and 0.12 respectively. A paired t-test comparing M_{Flex} and $M_{Flex}^{Continuous}$ rejected the null hypothesis that $E[u_{M_{Flex}} - u_{M_{Flex}^{Continuous}}] \leq 0$ with p-value numerically 0. Hence, cross validation decisively favored the correct heterogeneity model M_{Flex}.

In Sim_2, we generated data from a model we will call $M_{Norm}^{Continuous}$, which is identical to M_{Norm} but without a discrete component:

$$Y_{z=0,i} \sim N(\mu_{z=0,i}, 1)$$

$$Y_{z=1,i} \sim N(\mu_{z=1,i}, 1)$$

$$\mu^i_{z=0} = X_i \cdot (-2, -1, 0, 1, 2)$$

$$\Delta_i = X_i \cdot (-2, -1, 0, 1, 2)$$

$$\mu^i_{z=1} = \mu^i_{z=0} + \Delta_i$$

Again, we simulated a clinical trial with 500 patients in each arm. X again consisted of 5 predictor variables generated from a standard normal distribution. We then fit the same three models to this data that we fit to Sim_1. M_{Flex} exhibited the identifiability issues discussed in the previous section that can arise when there are no latent classes in the true data generating process and the error distributions are correctly (over-)specified. Different MCMC chains got stuck at very high or low values of α_G, but all chains converged to 0 for β_G and the correct values for β_Δ. The LOO-CV estimated expected posterior predictive densities for M_{Flex}, $M_{Flex}^{Continuous}$, and $M_{Flex}^{Constant}$ were .2799, .2801, and 0.133 respectively. The p-value from a paired t test comparing the samples from M_{Flex} and $M_{Flex}^{Continuous}$ was 0.001. So cross validation selected the simplest correct model $M_{Flex}^{Constant}$.

1.3.1 The Importance of Flexible Error Distributions

We simulated data from a model similar to M_{Norm} but with highly skewed error distributions:

$$
\begin{aligned}
Y^i_{z=0} &\sim Gamma(\mu^i_{z=0}, shape_0, scale_0) \\
Y^i_{z=1} &\sim Gamma(\mu^i_{z=1}, shape_1, scale_1) \\
G_i &\sim Bernoulli(p_i) \\
logit(p_i) &= -1 + X_i \cdot (-2, -1, 0, 1, 2) \\
\mu^i_{z=0} &= 5\mathbb{1}\{G_i = 2\} + X_i \cdot (-2, -1, 0, 1, 2) \\
\Delta_i &= 5\mathbb{1}\{G_i = 1\} + 15\mathbb{1}\{G_i = 2\} + X_i \cdot (-2, -1, 0, 1, 2) \\
\mu^i_{z=1} &= \mu^i_{z=0} + \Delta_i
\end{aligned}
\tag{1.11}
$$

X consisted of 5 predictor variables generated from a standard normal distribution. $Gamma(\mu, shape, rate)$ denotes a Gamma distribution shifted to have mean μ. We chose $shape_0 = shape_1 = 1$ and $scale_0 = scale_1 = 10$ so that the error distributions were highly skewed as in the figure below.

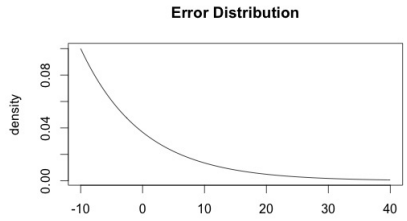

Figure 1.3: Skewed Error Distribution

We fit the models M_{Norm} and M_{Flex} described in the previous section to the data simulated from the above process. The two models only differed in their error distributions and were correctly specified in all other respects. Figure 4 compares the models' estimates of certain parameters of interest. We see that the M_{Norm}

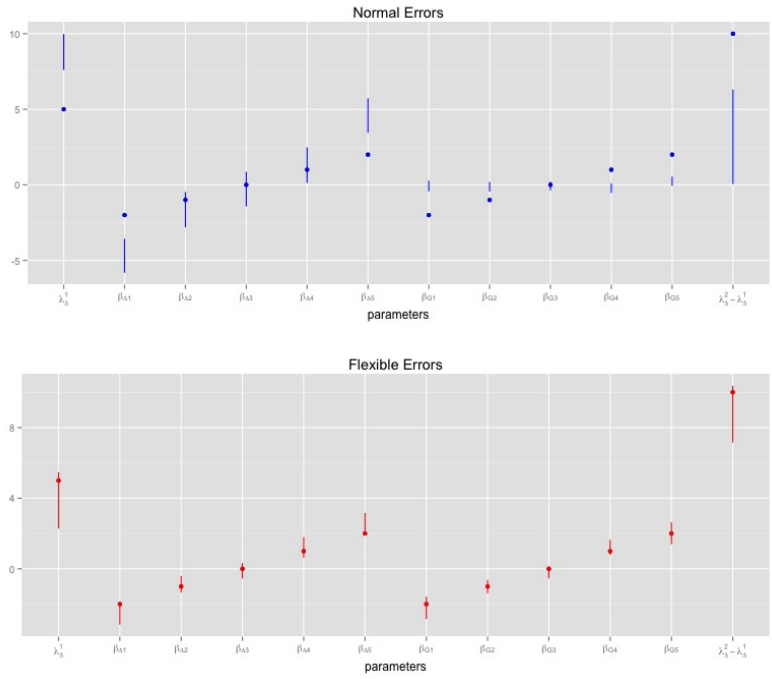

Figure 1.4: Comparison of M_{Norm} and M_{Flex}

estimates are off target for some parameters, including the $\lambda_\Delta^2 - \lambda_\Delta^1$ parameter that represents the magnitude of discrete heterogeneity between latent classes. The M_{Flex} estimates are fairly accurate for all parameters. These results illustrate sensitivity to misspecification of the error distribution and reassure us that the strategy of employing a flexible (mixture of normals) error distribution is sufficient to handle the problem.

To illustrate model comparison in this setting, we also consider $M_{Flex}^{Continuous}$. We use LOO-CV to compare $M_{Flex}^{Continuous}$ to M_{Flex}, which we know to be the superior model. In our simulated example, the LOO-CV estimated expected posterior predictive density of M_{Flex} was .04 compared to .03 for $M_{Flex}^{Continuous}$. A paired t-test comparing the samples of LOO-CV ppd's from the two models rejected the

null hypothesis that $E[u_{M_{Flex}} - u_{M_{Flex}^{Continuous}}] \leq 0$ with p-value numerically 0.

1.4 Re-analysis of Data from the ACTG 320 Clinical Trial

The ACTG 320 trial compared two AIDS treatments–a combination of indinavir, zidovudine, and lamivudine versus just zidovudine and lamivudine. Following (Shen and He, 2015) who themselves follow (Hammer et al., 1997) and (Zhao et al., 2013), we take change in CD4 count at the 24th week of treatment as the response variable, exclude patients with missing outcome values or extreme CD4 counts, and ignore any bias that we may induce by these exclusions. We are left with a dataset of 800 patients. A summary of the data is included in the Appendix.

Before fitting any models, we test the null hypothesis of a constant treatment effect using Rosenbaum's covariance adjustment test (Rosenbaum, 2002). The test produces a p value that is approximately 0, so we are quite certain that there is heterogeneity. The question remains whether it is related to observed covariates and whether we can effectively model it.

We fit multiple models and compare them using LOO-CV with posterior predictive density as the utility function. In some models, we use just the 3 covariates that Shen and He considered (baseline CD4, baseline RNA, and age), while in

others we take advantage of regularization to include the 9 other variables that were available. M_1 is Shen and He's model with two latent classes, a common normal error distribution for all patients, no continuous effect modification, and just 3 covariates. M_2 is a constant effect model with separate flexible error distributions for each treatment group. Note that 'constant effect' is a misnomer, since the distinct error distributions for the two treatment arms allow for heterogeneity, just not associated with the covariates. M_3 is the same as M_1 but with separate flexible error distributions for each treatment group. M_4 is the same as M_3 but includes all 12 covariates. M_5 is a continuous effect modification model with separate flexible error distributions for each treatment group and only 3 covariates. M_6 is the same as M_5 but includes all 12 covariates. M_7 is a 2 latent class mixture model with continuous effect modification and flexible error distributions and only 3 covariates. M_8 is the same as M_7 but includes all 12 available covariates. M_9 is the same as M_8 except that it has 3 latent classes instead of 2. All continuous effect modification was specified as linear, and all latent class membership models were specified as logistic or, in the case of M_9, multinomial logistic. The table at the top of Figure 5 summarizes key attributes of the models. A summary of parameter estimates from select models is in the Appendix.

Figure 5 depicts the LOO-CV estimated expected posterior predictive densities of each model. The first thing that jumps out in this plot is that M_1, which is Shen and He's model with a normal error distribution, performs far worse than the other models which all use flexible error distributions. This is not necessarily meaningful, however. Our parameters of interest do not govern the error distri-

	M1	M2	M3	M4	M5	M6	M7	M8	M9
Flexible Error Distributions		✔	✔	✔	✔	✔	✔	✔	✔
Continuous Heterogeneity					✔	✔	✔	✔	✔
Discrete Classes	2	0	2	2	0	0	2	2	3
# Covariates	3	0	3	12	3	12	3	12	12

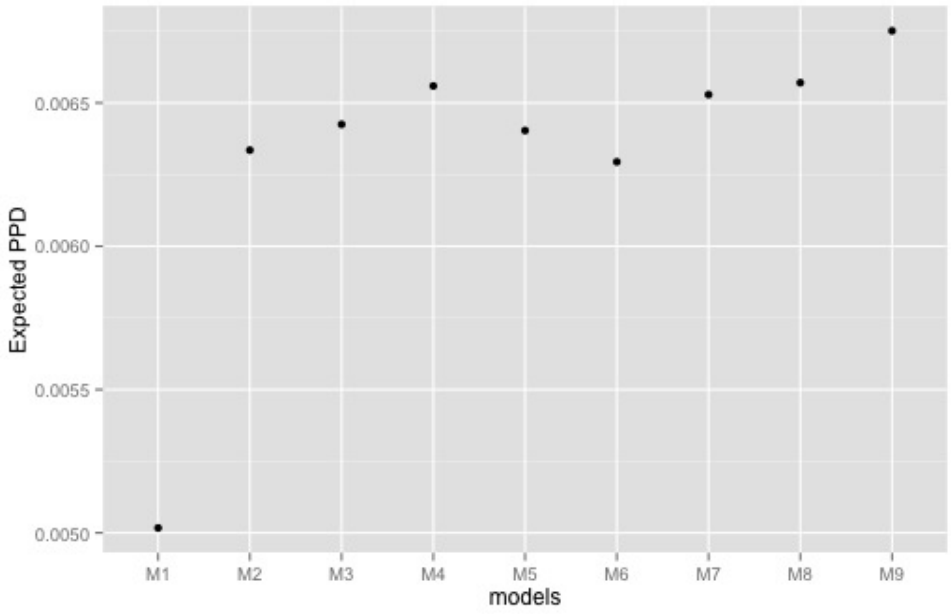

Figure 1.5: LOO-CV estimated expected posterior predictive densities of each model

bution, so the utility could be rewarding these models purely for better modeling an aspect of the data that is not important to us. But comparing the parameter estimates of M_3 (which is Shen and He's model with flexible error distributions) to M_1, we see that there are substantive differences. The estimated difference in treatment effects between latent classes is significantly smaller in M_3 than in M_1. Observing that the residuals are highly skewed and recalling the lessons learned from the simulation in Section 3.1, we suspect that the misspecified error distribution of M_1 biased the estimates of parameters of interest. However, it is still true that much of the difference in expected utility could be due to error distribution alone.

Next we turn our attention to the models with flexible error distributions. The most complex model, M_9, has the highest expected utility. When comparing models, we note that the estimated utilities are the sample means of the LOO-CV posterior predictive densities of the 800 observations. Since the samples of LOO-CV ppds produced by each model are based on the same observations, they are dependent and can be compared using classical methods for dependent samples such as paired t-tests or Wilcoxon signed rank tests. Paired t-tests indicate that the sample mean LOO-CV utility for M_9 is statistically significantly greater than the sample mean LOO-CV utilities of every other model. We can obtain a conservative p-value for the null hypothesis that M_9 has the highest true expected utility of all models considered by taking the p-value of the comparison between M_9 and the next best model (M_8) and adjusting for multiple comparisons using Holm's method. The paired t-test comparing M_9 to M_8 had p-value .006, and adjusting for the other 7 similar comparisons we might have made (i.e. testing

whether models 2 through 8 were the best) yields $p \approx .04$. That is, the probability of M_9 having such a superior estimated utility due to sampling variation alone if any of the other models had true expected utilities as good as M_9's is less than .04 ('less than' because our p-value is conservative). This indicates strong but not necessarily overwhelming support for M_9, so we would not completely dismiss other models with similar utilities such as M_8, M_7, and M_4. We definitely prefer M_9 but would take its implications with a grain of salt if they contradicted one of the other models with fairly similar utility.

We now take a closer look at what the models said about heterogeneity. Where comparisons between models could be made, the models were generally in agreement. First, every model found that heterogeneity was associated with the covariates (apart from the constant model $M2$, obviously). Of the three covariates that were included in every model (except M_2), baseline CD4 count and RNA levels, which we will denote $cd4_0$ and rna_0, were unanimously positively associated with treatment effect after adjusting for other covariates. The models that contained all 12 covariates also agreed that weight was positively associated and prior zidovudine exposure negatively associated with treatment effect adjusting for other covariates. (We will omit 'adjusting for other covariates' for the remainder of this discussion, but it should be understood that all associations might depend on which other covariates were included in the model.)

Every model that contained both continuous and discrete heterogeneity components (M_9, M_8, and M_7) attributed most covariate associated heterogeneity to discrete differences between latent classes. Every model that included latent classes

agreed that there was a substantial difference of about 60-80 CD4 count between the highest treatment effect class and the lowest. The three class model, M_9, also included a middle class with estimated treatment effect approximately 10 higher than in the lowest class. The treatment effect in the low class for an average patient was about 45-55 CD4 count in all models. All latent class models agreed that $cd4_0$ and rna_0 were positively associated with membership in the highest class. The models with 12 covariates also agreed that weight was positively associated and prior zidovudine negatively associated with membership in the highest class. In the two class models (M_3,M_4,M_7, and M_8), strength of association with class membership is easily discerned from the posterior distributions of the relevant logistic regression coefficients from the β_G parameter. In the preferred three class model, in which class is determined by a multinomial logistic regression, the relationship between β_G and the nature of the association is more subtle. Figure 6 compares the M_9 posterior distributions of probability of membership in the highest effect class for three hypothetical patients–one with the maximum observed value of $cd4_0$, one with the median observed value of $cd4_0$, and one with the minimum observed value of $cd4_0$. All three hypothetical patients were assigned median values for all other covariates. Comparing the high and medium patients, we see that the high $cd4_0$ value makes low probabilities of membership in the highest treatment effect class less likely but does not alter the mode. For very low values of $cd4_0$, membership in the high treatment effect class is virtually impossible. So M_9 appears to pick up on a nonlinear aspect to the association between $cd4_0$ and highest treatment effect class membership probability. Low values of $cd4_0$ are also strongly associated with membership in the middle class.

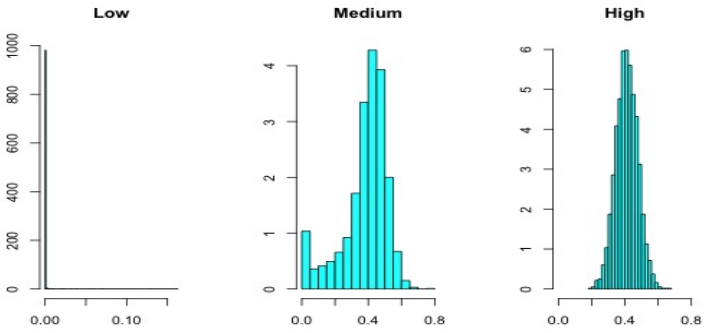

Figure 1.6: Posterior predictive distributions from model M_9 of probability of membership in the highest treatment effect class for hypothetical patients with low, median, and high $cd4_0$ values

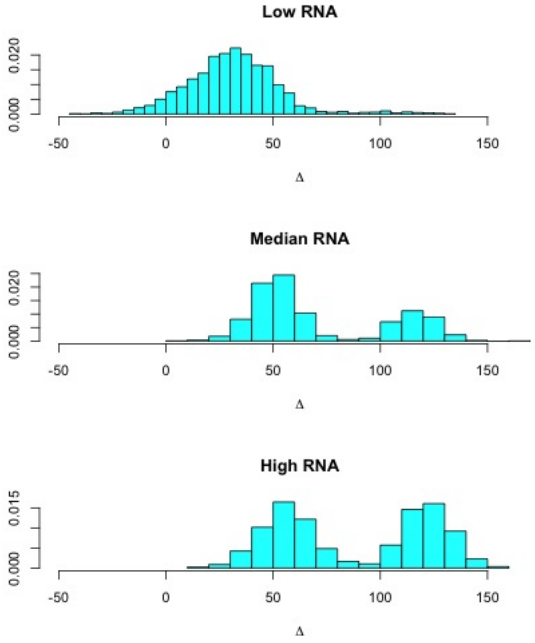

Figure 1.7: Posterior predictive distributions from model M_9 of the Δ parameter for hypothetical patients with low, median, and high rna_0 values.

The models that contained both continuous and discrete heterogeneity compo-

nents (M_9, M_8, and M_7) also all detected a possible moderate linear association with rna_0 and no other significant linear associations. So our preferred model M_9 and all the other credible models together imply mostly discrete heterogeneity associated with $cd4_0$, rna_0, weight, and zidovudine exposure along with possible modest continuous effect modification by rna_0. The strong performance of M_9 might be attributed to the more flexible relationships it allows between covariates and high effect class membership. A more thorough analysis would explore models with nonlinear regressions, more than 3 latent classes, interactions among covariates, and distinct error distributions for each latent class instead of just for each treatment group.

As a last step, we performed some basic posterior predictive checks (Gelman et al., 1996) to affirm that M_9 not only outperforms the other candidates but also fits the data reasonably well. First, Figure 8 compares a histogram of the outcomes from the real trial to a histogram of fake outcomes that were simulated from the posterior mean values of the parameters from M_9 and the observed covariate values. The distributions are remarkably similar. We then simulated 1000 fake data sets from the posterior predictive distribution of M_9, computed summary statistics of each fake data set, and checked whether the corresponding summary statistics of the true data fall within the range of the simulations. The exercise is summarized in figure 9 below, in which the red dots indicate the summary statistic values in the real data. The model appears to fit well by these criteria.

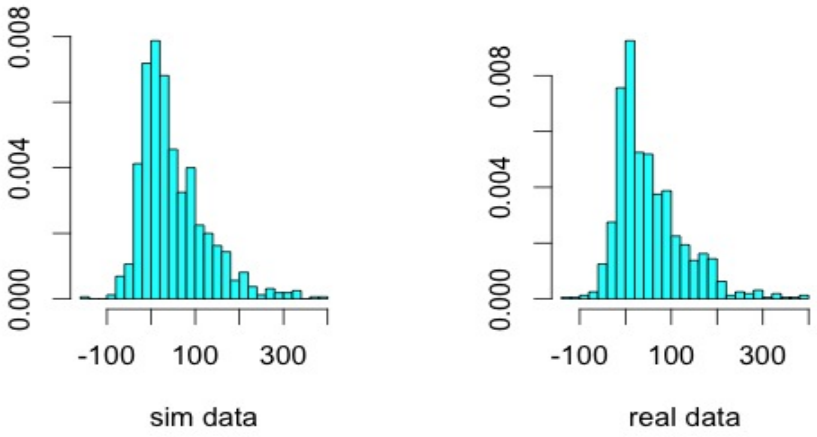

Figure 1.8: The distribution of the outcome variable in the trial (right) and a draw from the M_9 posterior predictive distribution of the outcome variable (left)

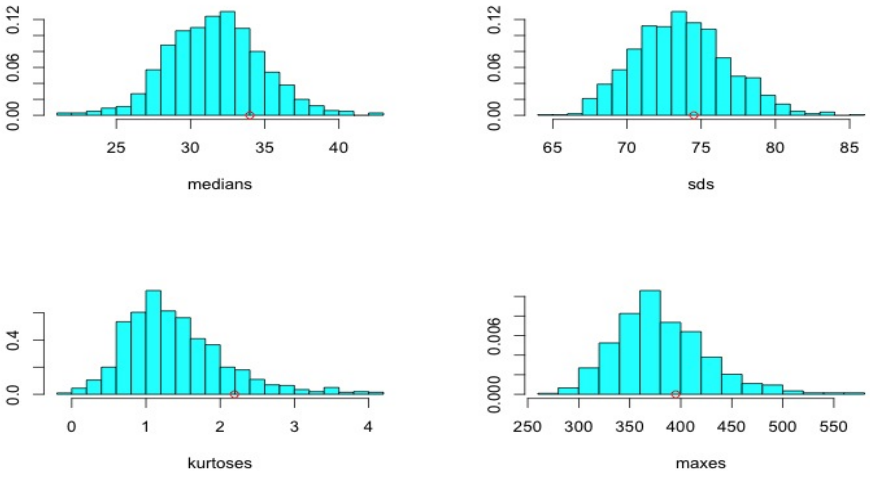

Figure 1.9: Posterior Predictive Checks of M_9

1.5 The Oregon Health Insurance Experiment (OHIE) Study

1.5.1 Description of the study and the data

In 2008, Oregon instituted a lottery to determine who could enroll in a new Medicaid program with limited openings. The randomness of the lottery induced a natural experiment that has allowed researchers to explore various public health and economic effects of Medicaid (Taubman et al., 2014). One important health economic question was what effect if any Medicaid might have on emergency department (ED) utilization. The intuitive and naive guess would be that health insurance would increase utilization by decreasing cost. However, many experts had predicted that expanding health insurance would actually decrease ED utilization for two main reasons. First, uninsured patients sometimes go to EDs for problems that could be addressed in a primary care setting because, unlike primary care physicians, EDs cannot turn patients away for being unable to afford treatment. Second, assuming Medicaid coverage would increase primary care utilization, regular monitoring of chronic conditions at primary care visits might prevent flareups that necessitate trips to the ED.

Taubman et al. addressed this question by looking at ED utilization among the 24,000 lottery participants who lived in Portland. They matched these lottery participants to medical records from 12 hospitals that accounted for almost all ED visits for Portland residents over the period of the study. Unfortunately, because many lottery winners did not go on to actually enroll in Medicaid and some lot-

tery losers managed to enroll through other channels, Taubman et al. could not simply compare lottery winners to losers to directly estimate the average causal effect (ACE) of Medicaid coverage. In these situations, the best one can do is to estimate the 'Complier Average Causal Effect' (CACE) using an instrumental variable analysis. The CACE is the average causal effect of Medicaid coverage on those lottery participants who would enroll in Medicaid if and only if they won the lottery (Imbens and Rubin, 1996; Rubin and Frangakis, 2002), i.e. the compliers. Taubman et al estimated the CACE to be positive with high confidence, supporting the naive and intuitive prediction that Medicaid coverage would increase ED utilization.

Restricting their data to approximately 10,000 lottery participants who filled out a survey containing questions pertaining to pre-treatment covariates, Taubman et al. explored heterogeneity in the CACE by performing both pre-registered and post hoc subgroup comparisons. They discovered several possible disparities in treatment effect (e.g. between smokers and non-smokers and between people with and without a prior serious chronic disease) but did not adjust either the pre-registered or the post-hoc comparisons for multiple testing. See the Appendix for tables summarizing the data and covariates.

1.5.2 Results of application to OHIE

We applied model M_{IV} defined in Section 2 to the OHIE data. In the context of the OHIE, the definitions of the principal strata are as follows: always takers would enroll in Medicaid regardless of whether they won the lottery; never takers

would not enroll in Medicaid regardless of whether they won the lottery; compliers would enroll if they won the lottery and not enroll if they lost; and defiers would enroll if they lost and not enroll if they won. We make the common assumption that there are no defiers. We get to observe the principal strata of some subjects, but other subjects' principal strata are latent. Lottery winners who don't enroll in Medicaid are definitely never takers, and lottery losers who do enroll are definitely always takers. But winners who enroll could either be compliers or always takers, and losers who do not enroll could either be compliers or never takers. G_i takes one value for never takers, one value for always takers, and one value for each treatment effect class for compliers to allow for discrete heterogeneity in the CACE. The model specified two treatment effect classes within the subgroup of compliers. Because we assume that the instrument only affects the outcome through the treatment, the instrument effect (represented by Δ_i in Figure 1) must be 0 whenever latent class G_i indicates a never-taker or always-taker.

When we included all recorded baseline covariates and used hierarchical shrinkage priors, we did not find evidence that treatment effect among compliers was associated with any of the recorded covariates. (Or, assuming that all covariates are associated with treatment effect at least a little bit, we did not find strong evidence of the direction of any of the associations.) The posterior distributions of all components of β_G and β_Δ were centered near zero with substantial probability mass on either side. A summary of the results is in the Appendix.

1.6 Conclusion

We have illustrated a general Bayesian framework for modeling treatment effect heterogeneity in experiments with non-categorical outcomes. Our modeling approach incorporates latent class mixture components to capture discrete heterogeneity and regression interaction terms to capture continuous heterogeneity. Flexible error distributions allow robust posterior inference on parameters of interest. Hierarchical shrinkage priors on relevant parameters address multiple comparisons concerns. Leave-one-out cross validation estimates of expected posterior predictive density obtained through importance sampling, together with posterior predictive checks, provide a convenient method for model selection and evaluation.

Simulated and real examples demonstrate the utility of this framework and the importance of its various features. The method provides convincing evidence that the heterogeneity in the the ACTG HIV trial is truly discrete and characterizes potential subgroups in terms of baseline covariates. Parameter estimates differ substantially from a prior analysis using a similar method (though the subjective interpretation of the output remains the same) as a result of using flexible error distributions, the importance of which is illustrated in simulations. In the IV analysis of the OHIE data, shrinkage priors serve their purpose and prevent premature identification of heterogeneities that may be due to multiple comparisons.

We see four immediate opportunities for future work. First, it should be relatively straightforward to develop implementations of this approach for other specialized outcome models, in particular for survival analyses. Second, if one could obtain

stable estimates of Bayes factors, possibly using the method of (Chib and Jeliazkov, 2001), more formal methods for model comparison with certain desirable properties would be available, and model averaged estimates of some relevant quantities could be computed. Third, variational Bayes approximations to the posteriors of these models would enable applications to experiments with very large numbers of covariates. And finally, nonparametric implementations could mitigate concerns about model misspecification.

1.7 Bibliography

Foster, Jared C., Jeremy MG Taylor, and Stephen J. Ruberg. "Subgroup identification from randomized clinical trial data." Statistics in medicine 30.24 (2011): 2867-2880.

Su, Xiaogang, et al. "Subgroup analysis via recursive partitioning." The Journal of Machine Learning Research 10 (2009): 141-158.

Rothwell, Peter M. "Subgroup analysis in randomised controlled trials: importance, indications, and interpretation." The Lancet 365.9454 (2005): 176-186.

Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." The Annals of Applied Statistics 7.1 (2013): 443-470.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. "Why we (usually) don't have to worry about multiple comparisons." Journal of Research on Educational Effectiveness 5.2 (2012): 189-211.

Titterington, D. Michael, Adrian FM Smith, and Udi E. Makov. Statistical analysis of finite mixture distributions. Vol. 7. New York: Wiley, 1985.

Sobel, Michael E., and Bengt Muthen. "Compliance Mixture Modelling with a Zero?Effect Complier Class and Missing Data." Biometrics 68.4 (2012): 1037-1045.

Rosenbaum, Paul R. "Covariance adjustment in randomized experiments and observational studies." Statistical Science 17.3 (2002): 286-327.

McLachlan, Geoffrey, and David Peel. Finite mixture models. John Wiley & Sons, 2004.

KANG, J., SU, X., HITSMAN, B., LIU, K. and LLOYD-JONES, D. (2012). Tree-structured analysis of treatment effects with large observational data. J. Appl. Stat. 39 513?529. MR2880431

M. Qian and S.A. Murphy (2011). Performance Guarantees for Individualized Treatment Rules. Annals of Statistics, Supplement. Apr 1;39(2):1180-1210. PMC3110016

Zhang, Baqun, et al. "A robust method for estimating optimal treatment regimes." Biometrics 68.4 (2012): 1010-1018.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107(499), 1106-1118.

Holm, Sture. "A simple sequentially rejective multiple test procedure." Scandinavian journal of statistics (1979): 65-70.

Dawid, Alexander Philip, and Monica Musio. "Theory and applications of proper scoring rules." Metron 72.2 (2014): 169-183.

Gelman, Andrew, Xiao-Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies." Statistica sinica 6.4 (1996): 733-760.

Chib, Siddhartha. "Analysis of treatment response data without the joint distribution of potential outcomes." Journal of Econometrics 140.2 (2007): 401-412.

Sarah Taubman, Heidi Allen, Bill Wright, Katherine Baicker, Amy Finkelstein, and the Oregon Health Study Group, "Medicaid Increases Emergency Department Use: Evidence from Oregon's Health Insurance Experiment", Science, 2014 Jan 17; 343(6168): 263-268

1.8 Appendix A: ACTG Trial

1.8.1 Data Summary

Table 1.1: ACTG Data Summary

	Min.	1st.Qu.	Median	Mean	3rd.Qu.
Female	0.00	0.00	0.00	0.16	0.00
Hemophilia	0.00	0.00	0.00	0.03	0.00
Weight	-4.21	-0.64	-0.07	-0.00	0.57
Karnofsky	-2.72	-0.15	-0.15	0.00	1.14
Prior_ZDV	-0.86	-0.69	-0.31	0.00	0.35
Age	-2.60	-0.68	-0.12	-0.00	0.58
CD4_0	-3.96	-0.65	0.26	-0.00	0.83
RNA_0	-4.85	-0.47	0.15	-0.00	0.68
IV_now	0.00	0.00	0.00	0.00	0.00
IV_past	0.00	0.00	0.00	0.14	0.00
Black or Hisp	0.00	0.00	0.00	0.44	1.00
Other_race	0.00	0.00	0.00	0.01	0.00

The Karnofsky score indicates the severity of the disease. Prior_ZDV indicates whether the patient had used similar drugs to those in the trial in the past. CD4_0 is baseline CD4 count, RNA_0 is baseline levels of HIV RNA in the blood. IV_now and IV_past indicate whether the patient is a current or past IV drug user, respectively. Below is a histogram of the outcome variable over all patients in the trial:

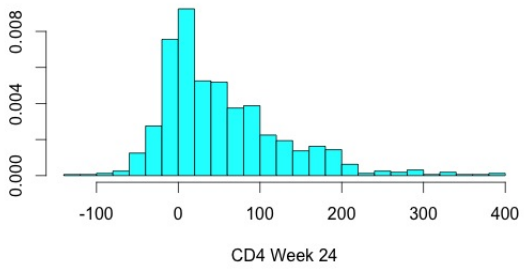

Figure 1.10: Histogram of Week 24 CD4 Count

1.8.2 Parameter Estimates For Select Models of ACTG Trial Data

M_9 Posterior Summary for Relevant Parameters:

	Median	Mean	SD	.05	.95
α_G^2	-2.56	-2.90	4.01	-10.10	3.14
α_G^3	-0.13	-0.13	0.29	-0.58	0.37
$\beta_{G,1}^2$	-1.53	-2.10	4.26	-9.68	4.11
$\beta_{G,2}^2$	-2.74	-2.99	6.67	-13.99	8.13
$\beta_{G,3}^2$	-7.29	-7.35	5.12	-15.87	1.38
$\beta_{G,4}^2$	-2.43	-2.73	2.21	-6.74	0.19
$\beta_{G,5}^2$	-1.24	-1.44	2.59	-5.71	2.23
$\beta_{G,6}^2$	-0.22	0.04	2.52	-3.49	4.81
$\beta_{G,7}^2$	-10.85	-10.94	5.26	-19.97	-2.97
$\beta_{G,8}^2$	-4.44	-4.58	2.67	-9.28	-0.84
$\beta_{G,9}^2$	0.85	1.31	6.38	-8.78	12.62
$\beta_{G,10}^2$	2.91	3.55	4.66	-3.16	12.10
$\beta_{G,11}^2$	-3.01	-3.48	3.70	-10.00	1.91
$\beta_{G,12}^2$	1.10	1.54	6.48	-9.14	12.84
$\beta_{G,1}^3$	0.03	0.05	0.25	-0.33	0.52
$\beta_{G,2}^3$	-0.03	-0.05	0.32	-0.61	0.43
$\beta_{G,3}^3$	-0.13	-0.14	0.19	-0.46	0.14
$\beta_{G,4}^3$	0.00	-0.00	0.15	-0.25	0.23
$\beta_{G,5}^3$	-0.21	-0.22	0.17	-0.52	0.02
$\beta_{G,6}^3$	-0.04	-0.04	0.16	-0.31	0.21
$\beta_{G,7}^3$	-0.15	-0.18	0.26	-0.65	0.17
$\beta_{G,8}^3$	0.33	0.34	0.21	0.00	0.70
$\beta_{G,9}^3$	-0.01	-0.02	0.32	-0.55	0.48
$\beta_{G,10}^3$	-0.06	-0.08	0.26	-0.54	0.29
$\beta_{G,11}^3$	0.16	0.19	0.24	-0.16	0.61
$\beta_{G,12}^3$	0.00	0.01	0.31	-0.48	0.52
$\beta_{\Delta,1}$	0.27	0.61	3.99	-5.68	7.60
$\beta_{\Delta,2}$	-1.22	-2.28	4.92	-11.47	3.98
$\beta_{\Delta,3}$	1.50	1.97	2.92	-2.12	7.27
$\beta_{\Delta,4}$	0.10	0.29	2.69	-3.92	4.96
$\beta_{\Delta,5}$	0.44	0.62	2.71	-3.76	5.33
$\beta_{\Delta,6}$	0.94	1.25	2.62	-2.90	5.93
$\beta_{\Delta,7}$	1.33	1.95	3.22	-2.41	7.94
$\beta_{\Delta,8}$	4.63	4.95	3.68	-0.10	11.46
$\beta_{\Delta,9}$	0.10	0.24	4.72	-7.22	8.11
$\beta_{\Delta,10}$	-0.65	-1.14	3.97	-8.05	4.74
$\beta_{\Delta,11}$	-1.05	-1.62	3.70	-8.38	3.87
$\beta_{\Delta,12}$	-0.01	0.06	4.57	-7.40	7.48
λ_Δ^1	45.64	45.60	6.34	35.35	55.81
$\lambda_\Delta^2 - \lambda_\Delta^1$	8.98	10.61	7.88	1.04	26.03
$\lambda_\Delta^3 - \lambda_\Delta^2$	62.78	62.26	11.81	41.67	80.67
σ_C	4.78	4.78	1.61	1.18	7.25
σ_Δ	3.80	4.02	2.26	0.61	8.13
σ_G^1	6.24	6.03	2.42	1.85	9.59
σ_G^2	0.27	0.29	0.15	0.08	0.57

The multinomial logistic regression parameters α_G^j and $\beta_{G,k}^j$ are such that $P(G_i = 1|X_i) = \frac{1}{1+e^{\alpha_G^2+\beta_G^2\cdot X_i}+e^{\alpha_G^3+\beta_G^3\cdot X_i}}$ and for j equal to 2 or 3 $P(G_i = j|X_i) = \frac{e^{\alpha_G^j+\beta_G^j\cdot X_i}}{1+e^{\alpha_G^2+\beta_G^2\cdot X_i}+e^{\alpha_G^3+\beta_G^3\cdot X_i}}$. The parameter indices match the rows of the table in Appendix A.

M_8 **Posterior Summary for Relevant Parameters:**

	Median	Mean	SD	.05	.95
α_G	-1.30	-1.34	0.42	-2.10	-0.73
$\beta_{G,1}$	0.21	0.23	0.33	-0.28	0.82
$\beta_{G,2}$	-0.07	-0.08	0.47	-0.87	0.66
$\beta_{G,3}$	0.22	0.22	0.16	-0.04	0.48
$\beta_{G,4}$	0.01	0.00	0.17	-0.28	0.28
$\beta_{G,5}$	-0.28	-0.30	0.19	-0.63	-0.00
$\beta_{G,6}$	0.07	0.08	0.17	-0.20	0.36
$\beta_{G,7}$	0.64	0.67	0.28	0.26	1.17
$\beta_{G,8}$	0.59	0.59	0.21	0.25	0.94
$\beta_{G,9}$	-0.07	-0.08	0.51	-0.93	0.71
$\beta_{G,10}$	-0.16	-0.18	0.32	-0.74	0.31
$\beta_{G,11}$	0.31	0.33	0.29	-0.12	0.83
$\beta_{G,12}$	0.02	0.02	0.51	-0.80	0.86
$\beta_{\Delta,1}$	-0.11	-0.16	3.93	-6.66	6.39
$\beta_{\Delta,2}$	-1.10	-1.97	4.67	-10.70	4.36
$\beta_{\Delta,3}$	1.37	1.67	2.70	-2.30	6.47
$\beta_{\Delta,4}$	0.25	0.43	2.65	-3.86	5.03
$\beta_{\Delta,5}$	0.51	0.69	2.72	-3.75	5.35
$\beta_{\Delta,6}$	1.02	1.28	2.61	-2.73	5.86
$\beta_{\Delta,7}$	-0.03	-0.02	2.58	-4.24	4.34
$\beta_{\Delta,8}$	4.90	5.19	3.69	-0.02	11.70
$\beta_{\Delta,9}$	0.10	0.28	4.58	-7.07	8.06
$\beta_{\Delta,10}$	-0.81	-1.41	4.01	-8.74	4.37
$\beta_{\Delta,11}$	-1.18	-1.73	3.81	-8.68	3.76
$\beta_{\Delta,12}$	0.08	0.24	4.43	-6.84	7.64
λ^1_Δ	51.49	51.57	5.80	42.23	61.18
$\lambda^2_\Delta - \lambda^1_\Delta$	57.62	57.34	11.81	37.32	76.07
σ_C	2.08	2.26	1.01	1.02	4.18
σ_Δ	3.78	3.97	2.17	0.63	8.04
σ_G	0.45	0.48	0.20	0.23	0.84

Parameters correspond exactly to M_{Flex}.

M_1 **Posterior Summary for Relevant Parameters:**

	Median	Mean	SD	.05	.95
α_G	-1.32	-1.33	0.25	-1.75	-0.95
$\beta_{G,1}$	0.48	0.49	0.22	0.15	0.86
$\beta_{G,2}$	0.79	0.80	0.24	0.42	1.21
$\beta_{G,3}$	-0.14	-0.14	0.21	-0.50	0.18
λ^1_Δ	36.26	36.25	4.66	28.70	43.74
$\lambda^2_\Delta - \lambda^1_\Delta$	105.35	105.50	12.99	84.41	127.01

This is Shen and He's model. It is a special case of M_{Norm} without continuous heterogeneity or regularization.

1.9 Appendix B: OHIE

1.9.1 Data Summary

Table 1.2: OHIE Covariate Summary

	Min.	.25	Median	Mean	.75	Max.
English	0.00	0.00	0.00	0.10	0.00	1.00
Female	0.00	0.00	1.00	0.56	1.00	1.00
First_Day	0.00	0.00	0.00	0.09	0.00	1.00
Age	-1.68	-0.91	-0.05	0.00	0.81	1.92
SNAP	0.00	0.00	1.00	0.55	1.00	1.00
TANF	0.00	0.00	0.00	0.02	0.00	1.00
Prior_ED	0.00	0.00	0.00	0.33	1.00	1.00
Edu	0.00	1.00	1.00	0.80	1.00	1.00
Black	0.00	0.00	0.00	0.12	0.00	1.00
Hisp	0.00	0.00	0.00	0.18	0.00	1.00
Other_Race	0.00	0.00	0.00	0.15	0.00	1.00
Smoker	0.00	0.00	0.00	0.43	1.00	1.00
Prior_DX	0.00	0.00	1.00	0.56	1.00	1.00

English indicates whether the patient required instructions in a language other than English. First_Day indicates whether the patient signed up for the Medicaid lottery on the first possible day. Age was of course standardized. SNAP and TANF indicate whether the patient had ever enrolled in other state assistance programs. Prior_ED indicates whether the patient had visited the ED in the year prior to the lottery. Prior_DX indicates whether the patient had a serious chronic disease such as diabetes, asthma, or cancer. Below is a histogram of ED utilization of all lottery participants:

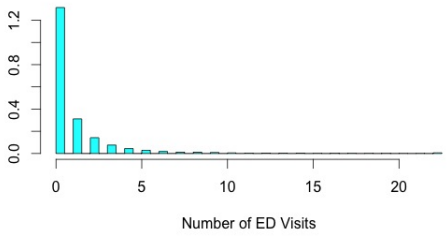

Figure 1.11: Histogram of ED Utilization

1.9.2 Posterior Summary of M_{IV} Applied to OHIE

	Median	Mean	SD	.025	.975
α_G^1	0.00	0.00	0.00	0.00	0.00
α_G^2	0.00	0.13	9.70	-18.25	19.63
α_G^3	-0.17	-0.21	9.30	-17.93	16.94
α_G^4	1.12	0.97	9.93	-18.89	19.51
$\beta_{G,1}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,1}^2$	-0.00	-0.06	2.39	-5.42	5.34
$\beta_{G,1}^3$	-0.01	-0.13	2.35	-5.48	4.97
$\beta_{G,1}^4$	-0.01	-0.12	2.38	-5.46	4.86
$\beta_{G,2}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,2}^2$	-0.00	-0.03	2.44	-5.47	5.34
$\beta_{G,2}^3$	0.00	-0.02	2.36	-5.57	5.08
$\beta_{G,2}^4$	-0.00	-0.02	2.39	-5.27	5.34
$\beta_{G,3}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,3}^2$	0.00	0.13	2.31	-4.78	5.62
$\beta_{G,3}^3$	0.00	0.03	2.33	-5.12	5.44
$\beta_{G,3}^4$	-0.00	0.04	2.36	-4.98	5.26
$\beta_{G,4}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,4}^2$	0.00	0.08	2.53	-5.64	5.97
$\beta_{G,4}^3$	0.00	0.01	2.36	-5.41	5.49
$\beta_{G,4}^4$	0.00	0.04	2.33	-4.90	5.40
$\beta_{G,5}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,5}^2$	0.00	-0.03	2.48	-5.48	4.92
$\beta_{G,5}^3$	-0.00	0.01	2.47	-5.24	5.57
$\beta_{G,5}^4$	0.00	-0.00	2.51	-5.25	5.54
$\beta_{G,6}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,6}^2$	0.00	0.12	2.37	-4.99	5.66
$\beta_{G,6}^3$	-0.00	0.01	2.28	-4.94	4.99
$\beta_{G,6}^4$	0.01	0.08	2.46	-5.13	5.69
$\beta_{G,7}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,7}^2$	-0.00	-0.08	2.30	-5.44	4.80
$\beta_{G,7}^3$	-0.01	-0.12	2.35	-5.36	5.10
$\beta_{G,7}^4$	-0.00	-0.02	2.24	-5.21	4.81
$\beta_{G,8}^1$	0.00	0.00	0.00	0.00	0.00
$\beta_{G,8}^2$	-0.01	-0.11	2.44	-5.71	5.00
$\beta_{G,8}^3$	0.00	0.07	2.32	-4.97	5.24
$\beta_{G,8}^4$	0.00	-0.02	2.40	-5.37	5.21
$\beta_{G,9}^1$	0.00	0.00	0.00	0.00	0.00

$\beta^2_{G,9}$	0.01	0.14	2.48	-4.94	5.79
$\beta^3_{G,9}$	-0.00	0.06	2.35	-4.99	5.51
$\beta^4_{G,9}$	-0.00	-0.03	2.40	-5.27	5.30
$\beta^1_{G,10}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{G,10}$	0.00	-0.06	2.36	-5.42	5.14
$\beta^3_{G,10}$	0.00	0.03	2.45	-5.06	5.94
$\beta^4_{G,10}$	-0.01	-0.05	2.31	-5.18	4.84
$\beta^1_{G,11}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{G,11}$	-0.00	-0.06	2.33	-5.30	5.02
$\beta^3_{G,11}$	0.01	0.11	2.42	-5.09	5.48
$\beta^4_{G,11}$	0.01	0.04	2.25	-4.87	4.94
$\beta^1_{G,12}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{G,12}$	0.01	0.10	2.45	-5.48	5.61
$\beta^3_{G,12}$	0.00	0.07	2.31	-5.07	5.10
$\beta^4_{G,12}$	0.01	0.12	2.35	-4.87	5.65
$\beta^1_{G,13}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{G,13}$	-0.02	-0.19	2.47	-5.71	5.15
$\beta^3_{G,13}$	0.01	0.15	2.36	-4.78	5.66
$\beta^4_{G,13}$	-0.00	-0.01	2.41	-5.14	5.22
$\beta^1_{\Delta,1}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,1}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,1}$	-0.00	-0.00	0.11	-0.24	0.24
$\beta^4_{\Delta,1}$	-0.00	-0.01	0.10	-0.24	0.18
$\beta^1_{\Delta,2}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,2}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,2}$	-0.01	-0.03	0.09	-0.25	0.13
$\beta^4_{\Delta,2}$	-0.03	-0.06	0.10	-0.34	0.09
$\beta^1_{\Delta,3}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,3}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,3}$	-0.02	-0.05	0.11	-0.34	0.12
$\beta^4_{\Delta,3}$	0.01	0.03	0.10	-0.15	0.29
$\beta^1_{\Delta,4}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,4}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,4}$	0.00	0.00	0.07	-0.14	0.17
$\beta^4_{\Delta,4}$	-0.02	-0.04	0.08	-0.22	0.08
$\beta^1_{\Delta,5}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,5}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,5}$	0.01	0.04	0.10	-0.13	0.30
$\beta^4_{\Delta,5}$	0.01	0.02	0.09	-0.16	0.23
$\beta^1_{\Delta,6}$	0.00	0.00	0.00	0.00	0.00

$\beta^2_{\Delta,6}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,6}$	0.01	0.02	0.11	-0.20	0.27
$\beta^4_{\Delta,6}$	-0.00	-0.00	0.11	-0.25	0.23
$\beta^1_{\Delta,7}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,7}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,7}$	0.00	0.01	0.10	-0.20	0.24
$\beta^4_{\Delta,7}$	-0.00	-0.01	0.09	-0.22	0.18
$\beta^1_{\Delta,8}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,8}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,8}$	-0.00	-0.00	0.09	-0.20	0.19
$\beta^4_{\Delta,8}$	-0.02	-0.05	0.11	-0.35	0.11
$\beta^1_{\Delta,9}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,9}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,9}$	-0.00	-0.01	0.09	-0.24	0.16
$\beta^4_{\Delta,9}$	0.00	0.01	0.10	-0.18	0.23
$\beta^1_{\Delta,10}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,10}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,10}$	-0.01	-0.03	0.11	-0.30	0.16
$\beta^4_{\Delta,10}$	0.02	0.04	0.10	-0.12	0.31
$\beta^1_{\Delta,11}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,11}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,11}$	-0.00	-0.02	0.10	-0.25	0.16
$\beta^4_{\Delta,11}$	-0.01	-0.02	0.10	-0.26	0.17
$\beta^1_{\Delta,12}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,12}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,12}$	0.02	0.05	0.11	-0.12	0.30
$\beta^4_{\Delta,12}$	0.01	0.03	0.10	-0.16	0.26
$\beta^1_{\Delta,13}$	0.00	0.00	0.00	0.00	0.00
$\beta^2_{\Delta,13}$	0.00	0.00	0.00	0.00	0.00
$\beta^3_{\Delta,13}$	-0.00	-0.01	0.09	-0.22	0.16
$\beta^4_{\Delta,13}$	0.01	0.03	0.10	-0.13	0.29
λ^1_Δ	0.00	0.00	0.00	0.00	0.00
λ^2_Δ	0.00	0.00	0.00	0.00	0.00
λ^3_Δ	0.11	0.09	0.19	-0.37	0.41
λ^4_Δ	0.47	0.48	0.17	0.19	0.86
r_1	0.48	0.48	0.03	0.43	0.54
r_2	0.84	0.84	0.08	0.70	1.00
r_3	5.79	6.00	1.94	2.85	9.72
r_4	0.43	0.44	0.06	0.33	0.57
σ_C	0.52	0.53	0.06	0.42	0.65

σ_G	1.92	1.97	1.39	0.01	4.73
σ_Δ	0.07	0.09	0.07	0.01	0.24

$G = 1$ corresponds to never takers, $G = 2$ corresponds to always takers, $G = 3$ corresponds to low treatment effect class compliers, and $G = 4$ corresponds to high treatment effect class compliers. The multinomial logistic regression parameters α_G^j and β_G^j are such that $P(G_i = j|X_i) = \frac{e^{\alpha_G^j + \beta_G^j \cdot X_i}}{1+e^{\alpha_G^2 + \beta_G^2 \cdot X_i} + e^{\alpha_G^3 + \beta_G^3 \cdot X_i}}$. Note that $\beta_G^1 = \beta_G^2 = \beta_\Delta^1 = \beta_\Delta^2 = 0$ because there is no treatment effect, and hence no treatment effect heterogeneity, for never takers and always takers. The λ_Δ and r parameters vary by group but could just be modeling the error distribution and are therefore not interpretable in terms of heterogeneity. The regression coefficient parameter indices match the rows of the covariate summary table.

Chapter 2

Ensembles of Granger Graphs (EGG) for Causal Discovery in High Dimensional Longitudinal Databases

2.1 Introduction

Health care decisions frequently depend on observational studies of health insurance claims databases. Setting aside important issues relating to data quality (e.g. Lewis et al., 2008; Harrold et al., 2007), such databases can be said to longitudinally track the occurrences of thousands of different types of health events in millions of patients. Given two health events $\mathbf{e}_1$ and $\mathbf{e}_2$ (e.g. sepsis and ischemic stroke), we consider the challenge of using a claims database to determine whether $\mathbf{e}_1$ causes $\mathbf{e}_2$.

Standard approaches to this problem estimate a confidence interval for the strength of association between occurrences of $\mathbf{e}_1$ and later occurrences of $\mathbf{e}_2$ adjusting for a set of possible confounding variables. A confidence interval excluding the value corresponding to no association constitutes evidence for a causal relationship. Many

combinations of study designs (e.g. cohort, self controlled case series, etc.) and statistical methods (e.g. propensity score matching, regression, etc.) are available to estimate the confidence interval for the covariate adjusted association, but for causal validity they all require an ignorability assumption to hold (Rosenbaum and Rubin, 1983). The ignorability assumption approximately states that there is no confounding conditional on the covariates included in the analysis. While it is impossible to test statistically whether ignorability holds for any particular study, past empirical experiments on data of this form (Madigan et al., 2014) suggest that it is the norm for failures of ignorability to be present and to lead to serious bias.

Valid causal discovery methods do exist, however, that do not assume ignorability. For example, building on earlier work by Richardson and Spirtes on Partial Ancestral Graphs (Richardson and Spirtes, 2002), Michael Eichler has developed a graphical model based framework for causal discovery in time series (Eichler, 2005; Eichler, 2007; Eichler, 2010; Eichler, 2012) in which it is sometimes possible to identify relationships between pairs of time series variables as causal or spurious without assuming that there are no unobserved confounders.

Eichler considered a situation in which we observe a subset of time series variables $\mathcal{O}$ from a larger stationary process $\mathcal{V}$. $\mathcal{V}$ is assumed to be sufficiently rich that, were we able to observe the full process, we could infer its causal structure from the conditional temporal associations between its variables. Note that not only do we not observe the variables in $\mathcal{V}\backslash\mathcal{O}$, but we may not even know what they are. For example, in our setting, $\mathcal{O}$ comprises a set of binary time series cor-

responding to the daily occurrence or non-occurrence of health events that appear in insurance claims records. We observe one realization of $\mathcal{O}$ for each patient. $\mathcal{V}$ would contain additional unobserved variables that are sufficient to causally explain the unfolding of patients' health histories–e.g. variables pertaining to diet or employment status. Clearly, $\mathcal{V}$ may be vast. Eichler's framework also requires that the data generating process for $\mathcal{V}$ belong to a wide class of processes that Eichler did not fully characterize but which includes, for example, all autoregressive models (Eichler, 2012).

We are particularly interested in $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{O}$. Eichler's insight was that certain configurations of conditional Granger non-causality (CGnC) relationships among $\mathcal{O}$ are only consistent with data generating processes for $\mathcal{V}$ in which certain pairs of observed variables are causally related (or spuriously associated). (A detailed definition of 'conditional Granger non-causality' will come later, but for now it suffices to think of it as a form of conditional independence for time series.) Thus, under certain assumptions about $\mathcal{V}$ mentioned above, there are certain sets of CGnC relations among $\mathcal{O}$ that, if they held, would imply $\mathbf{e}_1$ causes $\mathbf{e}_2$. Crucially, note that we need not assume that there are no latent common causes of variables in $\mathcal{O}$.

If the CGnC relations of a process $\mathcal{O}$ imply that a time directed association between a pair of variables is causal (or spurious), then we say that $\mathcal{O}$ *resolves* the association. Supposing that occurrences of $\mathbf{e}_1$ are associated with later occurrences of $\mathbf{e}_2$, an obvious strategy to evaluate whether $\mathbf{e}_1$ actually causes $\mathbf{e}_2$ is to search for a subset $\mathcal{O}$ that resolves their association. For large subsets $\mathcal{O}$, many CGnC tests are required to determine whether $\mathcal{O}$ resolves a given association. This is

problematic because (a) each CGnC test requires fitting a separate model; and (b) each individual CGnC test is prone to error, and an error in any one test may lead to faulty causal conclusions. Therefore, it seems wise to limit the search to smaller subsets. And because even causal conclusions derived from small subsets are not necessarily reliable, it seems wise to search for as many resolving subsets as possible to see if they produce a consensus.

We propose to select an ensemble of promising small subsets $\mathcal{O}_1, \ldots, \mathcal{O}_R$, each containing only three variables including the pair of interest. We will describe later how to find promising $\mathcal{O}_i$. For each $\mathcal{O}_i$, we learn its CGnC relations (which can be represented by a Granger causal graph), and we tally the conclusions of all the resolving subsets. If the resolving subsets produce a consensus that the association is causal (spurious), we regard that as strong evidence of causality (spurious association). If, as will often be the case, there is no consensus one way or the other, we consider this to be an absence of evidence but not necessarily evidence of absence. We call this procedure Ensemble of Granger Graphs, or EGG.

The hope is that a causal discovery method that does not rely on ignorability will have higher positive predictive value than standard methods that are sensitive to confounding. The price for ignoring ignorability is assuming restrictions on the full data generating process that are required for the theoretical validity of Eichler's framework. Whether this tradeoff is worthwhile, and whether EGG is robust to the violations of its assumptions that inevitably exist in practice, are empirical questions that we begin to address in this chapter. In Section 2, we explain Granger causal graphical models and dynamic Maximum Ancestral Graphs,

which are the tools Eichler developed for drawing limited causal conclusions from CGnC relations in the possible presence of latent confounding. In Section 3, we describe EGG in detail. In particular, we discuss our approach to CGnC testing. In Section 4, we present a simulation study in which EGG demonstrates excellent power and positive predictive value in the presence of unobserved confounding. In Section 5, we present results from applying EGG to a collection of health event pairs whose true (non-)causal relationships are known. EGG again demonstrates decent power and superior positive predictive value compared to a cohort method. We also apply EGG to two actual problems of interest in stroke research. In Section 6, we conclude.

2.2 Granger Causal Graphical Models

Here we briefly summarize Eichler's Granger graphical model framework for causal discovery from time series in the presence of confounding. A longer summary can be found in (Eichler, 2012) with further details in the references therein. In section 2.1, we consider the interpretation of Granger causal graphical models in the simpler context where we observe the full process and there are no unobserved confounders. Then we look at the case where we only observe a subset of the full process and latent confounding is possible. Throughout, we assume a structural equations framework that is more restrictive than strictly necessary.

2.2.1 Granger Causal Graphical Models For A Full Process

Suppose $\mathcal{V} = (V_1, \ldots, V_M)$ is a stationary multivariate time series process. Further suppose that the data generating process of $\mathcal{V}$ is defined by the structural equations

$$\begin{aligned} &V_i(t) = f_i(V_1^{t-1}, \ldots, V_M^{t-1}, U_i^t), \\ &i \in \{1, \ldots, M\} \end{aligned} \tag{2.1}$$

where $V_i(t)$ denotes the value of variable i at time t, V_j^{t-1} denotes the entire history of variable V_j through time $t-1$, and U_i^t is an error term representing the history of influential but unobserved variables not included in $\mathcal{V}$ through time t. We assume the underlying system is deterministic, though f_i is not necessarily known, but we can still speak about the probability distribution of $V_i(t)$ because U_i^t is unobserved.

Given such a system, we define two notions of causality and examine their relation to each other.

Definition 1 *Variable V_i* ***directly structurally causes*** *variable V_j if and only if f_j is not constant in V_i^{t-1}.*

The interpretation of direct structural causality is that one could in theory control the distribution of a variable by intervening on its direct structural causes. It is usually the goal of causal discovery to identify exactly such relationships.

Another popular concept of causality is conditional Granger (non-)causality. V_i is said to be Granger non-causal for V_j relative to $\mathcal{S} \subseteq \mathcal{V}$ if the history of V_i does not

help to predict the next value of V_j given the histories of $\mathcal{S}\backslash V_i$ and V_j. The special case where $\mathcal{S} = \mathcal{V}$ is often referred to plainly as Granger (non-)causality. A more formal definition is below:

Definition 2 *Given two variables $V_i, V_j \in \mathcal{V}$, we say that V_i is **Granger non-causal** for V_j relative to $\mathcal{S} \subseteq \mathcal{V}$ (denoted $V_i \nrightarrow_{GC} V_j|\mathcal{S}$) if $V_j(t) \perp\!\!\!\perp V_i^{t-1}|(\mathcal{S}\backslash V_i)^{t-1}, V_j^{t-1}$. Otherwise, V_i is Granger causal for V_j relative to $\mathcal{S}$ (denoted $V_i \rightarrow_{GC} V_j|\mathcal{S}$).*

Unlike direct structural causality, Granger causality does not have an immediate interpretation in terms of interventions, but it is easier to check empirically without access to the underlying structural equations. The question arises of when the two notions coincide. To address this question, we might consider two graphs representing relationships among the variables $\{V_1, \ldots, V_M\}$ in $\mathcal{V}$–the direct structural causality graph G_V^{DSC} with an edge $V_i \rightarrow V_j$ whenever V_i directly structurally causes V_j, and the Granger causality graph G_V^{GC} with an edge $V_i \rightarrow V_j$ whenever $V_i \rightarrow_{GC} V_j|\mathcal{V}$. Note that both graphs are directed and may contain cycles. We can then ask under what conditions the two graphs agree.

Suppose V_i does not directly structurally cause V_j, but there exists a latent common cause $L \notin \mathcal{V}$ of V_i and V_j such that $L(t-2)$ influences $V_i(t-1)$ and $V_j(t)$. Then $V_i(t-1)$ and $V_j(t)$ would be associated even after adjusting for all variables in $\mathcal{V}$, i.e. V_i would Granger cause V_j. So it is clear that for there to be any hope of Granger causality agreeing with direct structural causality, external common causes cannot be allowed. To enforce that no association between variables in $\mathcal{V}$ is due to a common cause not included in $\mathcal{V}$, we assume the following independence

condition involving the error terms:

$$\begin{aligned} U_i(t) &\perp\!\!\!\perp V_j^t | V_i^t, U_j^t \\ \forall i, j &\in \{1, \ldots, M\}. \end{aligned} \tag{2.2}$$

The absence of external common causes is the sense in which we can consider $\mathcal{V}$ to be a 'full' process. However, see (Robins and Richardson, 2013) for an explanation of how assumptions about independent errors can have subtle implications beyond the absence of external common causes.

It turns out that for processes $\mathcal{V}$ meeting our assumption (2), Granger causality implies direct structural causality (White and Lu, 2010).

Proposition 1 *(White and Lu, 2010) Suppose $V_i, V_j \in \mathcal{V}$ where $\mathcal{V}$ is generated by structural equations (1) and has error terms satisfying the independence assumption in (2). Then if V_j does not directly structurally cause V_i, $V_i(t) \perp\!\!\!\perp V_j^{t-1} | (\mathcal{V} \backslash V_j)^{t-1}$.*

By Proposition 1, every edge in G_V^{GC} must also be present in G_V^{DSC}. Eichler and Didilez (2010) proved a partial converse to Proposition 1 under further assumptions on $\mathcal{V}$. Eichler and Didilez make the following two technical assumptions:

The conditional distribution of $\mathcal{V}(t+1)$ given $\mathcal{V}^t$ is almost surely absolutely continuous with respect to a product measure on $\mathbb{R}^M$ and has positive density almost everywhere. (2.3)

For all subprocesses $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$, $\mathcal{A}$ and $\mathcal{B}$ are measurably separated conditional on $\mathcal{C}$.

(2.4)

These assumptions are satisfied by a wide class of time series models including all autoregressive models. Under these assumptions, Eichler and Didilez prove:

Proposition 2 *(Eichler and Didilez, 2010) For variables $V_i, V_j \in \mathcal{V}$,*
(A) if there is a causal effect of intervening in $V_i(t)$ on $V_j(t+1)$, then $V_i \to V_j \in G_V^{GC}$;
(B) if there is a causal effect of intervening in $V_i(t)$ on $V_j(t+h)$ for any h, then there is a directed path $V_i \to \cdots \to V_j$ in G_V^{GC}.

Proposition 2 is almost a converse to Proposition 1 but not quite. Proposition 2 is framed in terms of effects of interventions whereas Proposition 1 is framed in terms of direct structural causality. It is possible, for example, for variable V_1's value at time $t-2$ to directly structurally cause V_2 at time t but for the effect of intervening on $V_1(t-2)$ to be exactly canceled out by an indirect effect that runs through a third variable V_3's value at time $t-1$. Under the further assumption that no direct structural causal effects are exactly canceled out by indirect effects (often referred to as a $faithfulness$ assumption in the graphical model literature), Proposition 1 and Proposition 2 together imply that G_V^{DSC} and G_V^{GC} are identical.

To summarize, for a full process $\mathcal{V}$ that satisfies the faithfulness assumption and technical conditions that are not terribly restrictive, Granger causality corresponds to direct structural causality and the graph G_V^{GC} of Granger causality relations amongst the variables is identical to the graph G_V^{DSC} of direct structural causality

relations amongst the variables. Since it is feasible to evaluate Granger causality empirically, these results imply that it is a useful tool for causal discovery when we observe a full process. Of course, we rarely get to observe a full process. In the next subsection, we discuss what conclusions may be drawn if we observe only a subset of a full process.

2.2.2 Granger Causal Graphical Models For Processes That May Contain Latent Variables

Let $\mathcal{V}$ be a multivariate time series process satisfying all the assumptions from the previous section, and suppose that we only get to observe $\mathcal{O} \subset \mathcal{V}$. We may not even know how many variables are in $\mathcal{V} \backslash \mathcal{O}$.

Eichler showed that all *conditional* Granger non-causality relationships amongst variables in $\mathcal{V}$ (i.e. relations of the form $V_j \nrightarrow_{GC} V_i | \mathcal{S}$ for any $\mathcal{S} \subseteq \mathcal{V}$) can be derived from the structure of G_V^{GC} using a path criterion called m-separation. For details on m-separation, consult (Eichler, 2012). To graphically represent the conditional Granger non-causality relations among just the variables in $\mathcal{O}$, Eichler had to develop a new type of graph (called a dynamic Maximal Ancestral Graph or dMAG) with two additional arrow types and a new accompanying heuristic rule called m*-separation. A dMAG over variables $\mathcal{O} = \{O_1, ..., O_K\}$ consists of edges

of the form:

$$O_i \to O_j, \tag{2.5}$$

$$O_i \dashrightarrow O_j, \text{ and} \tag{2.6}$$

$$O_i - - - O_j \tag{2.7}$$

Again, see (Eichler, 2012) for more details. One necessary shortcoming of dMAGs is that, unlike Granger causal graphs for full processes, different dMAGs may represent the conditional Granger non-causality relations of the same process. The class of all dMAGs representing a set of CGnC relations is called a *Markov equivalence class*. We say that the set of all dMAGs representing the CGnC relations of a process $\mathcal{O}$ is the Markov equivalence class for $\mathcal{O}$. dMAGs have several properties that make them useful tools for causal discovery. First, all dMAGs in the same Markov equivalence class have the same skeleton, which means that they have dashed line edges like (7) in all the same places and they have arrows in all the same places and pointing in the same direction though the tails of the arrows may differ in type across the Markov equivalence class between (5) and (6). This skeleton can be learned empirically from data with conditional independence tests. Second, given G_V^{GC}, it is possible to construct a dMAG over $\mathcal{O}$ (call it $dMAG_{\mathcal{V}}(\mathcal{O})$) such that $O_i \to O_j \in dMAG_{\mathcal{V}}(\mathcal{O})$ if and only if $O_i \to \cdots \to O_j \in G_V^{GC}$. In other words, it is always possible to construct a dMAG over an observed subset that preserves the ancestral causal relationships from the full process in the form of solid arrows of type (5).

Thus, if we are presented with a set of variables $\mathcal{O}$ that we are willing to assume

is a subset of an unknown full process $\mathcal{V}_{TRUE}$ satisfying the assumptions from the previous section, we can use the properties of dMAGs to reason as follows. There exists a dMAG (which we call $dMAG_{V_{TRUE}}(\mathcal{O})$) representing the CGnC relations in $\mathcal{O}$ such that every arrow of type (5) in $dMAG_{V_{TRUE}}(\mathcal{O})$ corresponds to a true ancestral causal relationship in $\mathcal{V}_{TRUE}$. Therefore, if an arrow of type (5) appears in every dMAG in the Markov equivalence class for $\mathcal{O}$, it must in particular be present in $dMAG_{V_{TRUE}}(\mathcal{O})$ and indicate a true ancestral causal relationship. Similarly, if every dMAG in the Markov equivalence class for $\mathcal{O}$ contains an arrow of type (6), then in particular $dMAG_{V_{TRUE}}(\mathcal{O})$ must contain the arrow of type (6), which implies that the variable at the tail of the arrow is not an ancestral cause of the variable at the head. This reasoning is diagrammed in Figure 1.

The entire Markov equivalence class of $\mathcal{O}$ can in theory be learned empirically from data and the unanimous arrows systematically identified. However, this is unnecessary as Eichler described simple sufficient conditions for concluding that a $\rightarrow$ (or $\dashrightarrow$) is present in every dMAG in an equivalence class. Here we describe a simplified version of these conditions for the case where $\mathcal{O}$ contains three variables. Certain complications arise with four or more variables (Section 6 of Eichler, 2012), but in our proposed procedure we never fit graphs with more than three variables.

First, the skeleton for a three variable process can be learned as follows. If $O_i \rightarrow_{GC} O_j | \mathcal{O}$ and $O_i \rightarrow_{GC} O_j | \emptyset$, then the skeleton of the Markov equivalence class for $\mathcal{O}$ contains $O_i \cdots\!\blacktriangleright O_j$ (where '$\cdots\!\blacktriangleright$' denotes an arrow whose tail may be of type (5) or (6)).

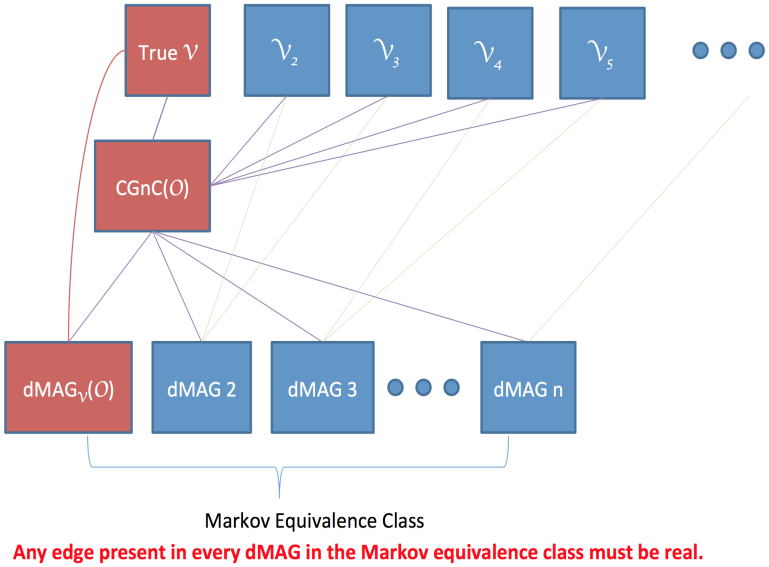

Figure 2.1: In the top row are the are infinitely many full processes that could have produced the observed CGnC relations $CGnC(\mathcal{O})$. Each of them has ancestral causal relations corresponding to the solid arrows in one of the n dMAGs in the Markov equivalence class representing $\mathcal{O}$ in the bottom row. $dMAG_V(\mathcal{O})$ has solid arrows corresponding to the ancestral causal relationships in the true full process $\mathcal{V}$. The lines in the diagram from the top row to the bottom row connect processes to their corresponding dMAGs. Any solid arrow present in every dMAG in the bottom row is in particular in $dMAG_{\mathcal{V}}(\mathcal{O})$ and therefore in True $\mathcal{V}$.

Proposition 3 *If a skeleton for* $\mathcal{O} = \{O_1, O_2, O_3\}$ *contains* $O_1 \cdots\!\blacktriangleright O_2 \cdots\!\blacktriangleright O_3$ *but not* $O_1 \cdots\!\blacktriangleright O_3$*, then it is possible to determine whether* O_2 *causes* O_3 *or the association is spurious. If* $O_1 \rightarrow_{GC} O_3|\emptyset$ *but* $O_1 \nrightarrow_{GC} O_3|O_2$*, then* $O_2 \rightarrow O_3$ *must be in every dMAG in the Markov equivalence class for* $\mathcal{O}$ *and therefore* O_2 *must cause* O_3*. If* $O_1 \nrightarrow_{GC} O_3|\emptyset$*, then* $O_2 \dashrightarrow O_3$ *must be in every dMAG in the Markov equivalence class for* $\mathcal{O}$ *and therefore the association between* O_2 *and* O_3 *must be spurious.*

The intuition behind this result is that causal effects pass temporal associations forward in time. Therefore, it should be necessary to condition on a cause to block the association between its precursor and its effect. Spurious associations do not carry associations forward in time, and it should not be necessary to condition on a spurious cause in order to block the association between its precursor and its spurious effect.

2.3 Ensembles of Granger Graphs (EGG)

We explain our proposed procedure in the context of a large longitudinal database of health histories, though it should be obvious how it would extend to other similar situations, e.g. simultaneous observations of large numbers of neural spike trains. Suppose there are P patients indexed by $p \in \{1, \ldots, P\}$, and the p^{th} patient is observed for N_p days. In each patient, the same M health events $\{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ are monitored. The p^{th} patient's history comprises a collection of binary time series $\{\mathbf{e}_1^p, \ldots, \mathbf{e}_M^p\}$ documenting the occurrence or non-occurrence of each health event on each day. $\mathbf{e}_i^p$ is a vector of dimension N_p such that $\mathbf{e}_i^p(t) = 1$ if event $\mathbf{e}_i$ occurred

on day t to patient p and 0 otherwise.

Suppose that in our P observations of $\{\mathbf{e}_1^p, \ldots, \mathbf{e}_M^p\}$, we have detected a large association between occurrences of health event $\mathbf{e}_i$ and later occurrences in the same patient of $\mathbf{e}_j$. For example, $\mathbf{e}_i$ may be ischemic stroke and $\mathbf{e}_j$ may be paralysis. To address the question, 'Does $\mathbf{e}_i$ actually cause $\mathbf{e}_j$?', we propose learning an ensemble of dMAGs containing $\mathbf{e}_i$ and $\mathbf{e}_j$. The procedure has three general steps:

Step 1. Select a set of candidate subprocesses $\{\mathcal{O}_1, \ldots, \mathcal{O}_R\}$, each containing $\mathbf{e}_i$ and $\mathbf{e}_j$, that are likely to resolve the association between $\mathbf{e}_i$ and $\mathbf{e}_j$.
Step 2. For each $\mathcal{O}_k$, conduct a series of conditional Granger causality tests to learn the skeleton of the Markov equivalence class for $\mathcal{O}_k$ and, if $\mathcal{O}_k$ resolves the association between $\mathbf{e}_i$ and $\mathbf{e}_j$, determine the causal implications of $\mathcal{O}_k$.
Step 3. Tally the results and summarize the evidence across subprocesses.

There are many options for how to perform each step of this general procedure. Below we discuss some of the considerations involved in each step and describe the choices we made in our experiments.

2.3.1 Step 1: Selecting Subprocesses

We prefer to use small subprocesses with just one variable in addition to $\mathbf{e}_i$ and $\mathbf{e}_j$. This is because the number of conditional Granger causality tests required for each process in Step 2 increases quickly with the number of variables. Therefore,

large processes will be both computationally expensive and unreliable. The reason more conditional Granger causality tests implies less reliability is that an error in any single test can alter the direction of evidence supplied by a process.

We want to choose subprocesses $\mathcal{O} = \{\mathbf{e}_h, \mathbf{e}_i, \mathbf{e}_j\}$ that are likely to resolve the association of interest between $\mathbf{e}_i$ and $\mathbf{e}_j$. Proposition 3 suggests that this amounts to choosing variables $\mathbf{e}_h$ that Granger cause $\mathbf{e}_i$ both conditional on the empty set and on $\mathbf{e}_j$. This way, $\mathbf{e}_h \cdots\!\blacktriangleright \mathbf{e}_i$ will be in the skeleton. Given that $\mathbf{e}_i$ and $\mathbf{e}_j$ were found to be associated in a prior adjusted analysis, $\mathbf{e}_i \cdots\!\blacktriangleright \mathbf{e}_j$ will also likely be in the skeleton. So by Proposition 3, all that's needed for $\mathcal{O}$ to resolve the association is for $\mathbf{e}_h$ to be Granger non-causal for $\mathbf{e}_j$ conditional on $\emptyset$ or $\mathbf{e}_i$.

We would further like to choose subprocesses that have low probability of resolving the association incorrectly due to statistical errors. There are two ways in which our procedure might go wrong for a given subprocess. $\mathbf{e}_i$ could be only spuriously associated with $\mathbf{e}_j$ but the subprocess could resolve it as causal (call this a type 1 error), or vice versa (a type 2 error).

Type 1 errors occur when we are able to detect that $\mathbf{e}_h \rightarrow_{GC} \mathbf{e}_j|\emptyset$ but not that $\mathbf{e}_h \rightarrow_{GC} \mathbf{e}_j|\mathbf{e}_i$. (See Proposition 3.) This is likely to happen when the association between $\mathbf{e}_h$ and $\mathbf{e}_i$ is very strong and there are not enough instances of $\mathbf{e}_h$ occurring without being followed by $\mathbf{e}_i$ to provide sufficient power to detect that $\mathbf{e}_h$ predicts $\mathbf{e}_j$ conditional on $\mathbf{e}_i$. We can reduce the chances of type 1 error by requiring that $\mathbf{e}_h$ has prevalence above some threshold in the data and is at least occasionally not followed by $\mathbf{e}_i$.

Type 2 errors occur when we fail to detect that $\mathbf{e}_h \rightarrow_{GC} \mathbf{e}_j|\emptyset$ even though $\mathbf{e}_i$ causes $\mathbf{e}_j$. This is likely to happen if the unadjusted association between $\mathbf{e}_h$ and $\mathbf{e}_i$ is weak or if the causal effect of $\mathbf{e}_i$ on $\mathbf{e}_j$ is weak. This is because the only association guaranteed to exist between $\mathbf{e}_h$ and $\mathbf{e}_j$ is the indirect combination of the association between $\mathbf{e}_h$ and $\mathbf{e}_i$ and the causal effect of $\mathbf{e}_i$ on $\mathbf{e}_j$. This indirect association will be weaker than either of its component parts. We of course cannot control the strength of the causal effect of $\mathbf{e}_i$ on $\mathbf{e}_j$, but we can require that the strength of the unadjusted association between $\mathbf{e}_h$ and $\mathbf{e}_i$ is above some minimum threshold. We can also again require that the prevalence of $\mathbf{e}_h$ is above a minimum threshold. Together, these requirements would ensure that there is sufficient power to detect an association between $\mathbf{e}_h$ and $\mathbf{e}_j$ as long as the causal effect of $\mathbf{e}_i$ on $\mathbf{e}_j$ is not extremely weak.

In practice, we collect promising candidates for the role of precursor variable $\mathbf{e}_h$ as follows. First, we decide the maximum number R of variables to select. (In our experiments in Section 5, we set $R = 50$.) We then define a measure of unadjusted strength of temporal association to be used for screening. In our experiments, we used a crude 'reverse incidence ratio' for this purpose. Specifically, we defined all days preceding an occurrence of $\mathbf{e}_i$ to be 'exposed' and all other days to be unexposed, then took

$$[(\#\mathbf{e}_h \text{ occurrences on exposed days})/(\#\text{ exposed days})]/$$
$$[(\#\mathbf{e}_h \text{ occurrences on unexposed days})/(\#\text{ unexposed days})]$$

to be the strength of temporal association between $\mathbf{e}_h$ and $\mathbf{e}_i$. This measure rewards events $\mathbf{e}_h$ that disproportionately precede $\mathbf{e}_i$. Next, we search the dataset for all variables that exceed some minimum allowable unadjusted strength of association with $\mathbf{e}_i$ (reverse incidence ratio > 2 in our experiments) and some minimum allowable population prevalence (.001 in our experiments). These minimum thresholds are chosen dependent on sample size to ensure sufficient power for relevant Granger causality tests. In our database, it was possible to quickly compute incidence ratios and prevalences for thousands of candidate variables using a few SQL commands. We then exclude any variables that only rarely occur without being followed by $\mathbf{e}_i$, and if there are more than R variables remaining we select the R with the highest incidence ratios.

2.3.2 Step 2: Conditional Granger Causality Testing

As described in Section 2.2, learning the skeleton of a Granger causal dMAG for process $\mathcal{O} = \{O_1, O_2, O_3\}$ and hopefully identifying one of its arrows consists of performing a series of conditional Granger causality tests. There has been work on nonparametric conditional Granger causality testing (Diks and Panchenko, 2005; Su and White, 2008; Dhamala et al., 2008), but existing methods are still slow and lack power, particularly for binary data. We therefore employ model based tests.

To test whether variable O_2 is Granger causal for variable O_3 conditional on variable S (where S could be O_1 or the empty set), we fit two models for O_3–one which

incorporates information from the histories of S, O_2, and O_3 (call this model $M_{S,2}$) and one which only incorporates information from the histories of S and O_3 (call this model M_S). We assign prior probability of $1/2$ to each model and compute each model's posterior probability given the data. If $P(M_{S,2}|Data) > 1 - \delta$ for some δ, we declare Granger causality. If $P(M_{S,2}|Data) < \delta$, we declare Granger non-causality.

We use simple piecewise constant conditional intensity models (PCIMs) for $M_{S,2}$ and M_S. See (Gunawardana et al, 2011) for a more complete discussion of PCIMs. A PCIM for the process $\mathcal{O}$ comprises a set of 'structures' $A = \{A_1, A_2, A_3\}$ and parameters λ. The structure A_i is a set of discrete 'states' that the history of $\mathcal{O}$ may satisfy at any given time. For example, a simple structure with states that incorporate the histories of all three variables might contain one state for each of the 8 possible combinations of binary indicators for past occurrences of each variable. Or structures could be richer, including states defined by how long ago each variable occurred or the number of times each variable has occurred. For each state $a \in A_i$, there is a corresponding conditional intensity λ_{ia} which is the probability of O_i occurring on any day satisfying a. Let x represent realizations of $\mathcal{O}$. Then the likelihood of the data x is

$$p(x|A, \lambda) = \prod_i \prod_{a \in A_i} \lambda_{ia}^{c_{ia}(x)} e^{-\lambda_{ia} d_{ia}(x)} \tag{2.8}$$

where $c_{ia}(x)$ denotes the number of times that O_i occurs in x under state a and d_{ia} denotes the total number of days that x is in state a.

We can put independent conjugate Gamma priors on the conditional intensity parameters.

$$p(\lambda_{ia}|\alpha_{ia},\beta_{ia}) = \frac{\beta_{ia}^{\alpha_{ia}}}{\Gamma(\alpha_{ia})}\lambda_{ia}^{\alpha_{ia}-1}e^{-\beta_{ia}\lambda_{ia}} \tag{2.9}$$

The conjugate posteriors for the conditional intensity parameters are then

$$p(\lambda_{ia}|\alpha_{ia},\beta_{ia},x) = p(\lambda_{ia}|\alpha_{ia}+c_{ia}(x),\beta_{ia}+d_{ia}(x)) \tag{2.10}$$

And, given independent priors on the λ parameters, the marginal likelihood is

$$\begin{aligned} p(x|A) &= \prod_i \prod_{a\in A_i} \gamma_{ia}(x); \\ \gamma_{ia}(x) &= \frac{\beta_{ia}^{\alpha_{ia}}}{\Gamma(\alpha_{ia})}\frac{\Gamma(\alpha_{ia}+c_{ia}(x))}{(\beta_{ia}+d_{ia}(x))^{\alpha_{ia}+c_{ia}(x)}} \end{aligned} \tag{2.11}$$

Returning to our example where we want to evaluate whether O_2 Granger causes O_3 relative to O_S, let $M_{S,2}$ and M_S be two PCIMs for $\mathcal{O}$ with the same structures A_1 and A_2 but different A_3. Because the marginal likelihood for a PCIM factors by variable, the Bayes factor comparing $M_{S,2}$ and M_S will be the ratio of the third factors in their respective marginal likelihoods, i.e.

$$BF(M_{S,2},M_S) = \prod_{a\in A_3^{S,2}} \gamma_{3a}(x) / \prod_{a\in A_3^S} \gamma_{3a}(x) \tag{2.12}$$

And the posterior probability of $M_{S,2}$ will be

$$P(M_{S,2}|x) = \frac{BF(M_{S,2},M_S)}{1+BF(M_{S,2},M_S)} \tag{2.13}$$

In our experiments and simulations, we set the hyper parameters α and β both equal to 1 for all variables and states, so that the Bayes factor is dominated by the likelihood. We used simple structures for $A_3^{S,2}$ and A_3^S. If $S = O_1$, A_3^S contains one state for each of the 4 possible combinations of binary indicators for past occurrences of O_1 and O_3, and $A_3^{S,2}$ contains one state for each of the 8 possible combinations of binary indicators for past occurrences of O_1, O_2 and O_3. If S is the empty set, A_3^S contains just two states: one corresponding to a past occurrence of O_3 and one corresponding to no past occurrence of O_3. And $A_3^{S,2}$ would contain the 4 states corresponding to the different combinations of past occurrences of O_2 and O_3. Of course, it is clear how to modify the above for any choice of variables to be the Granger cause, Granger effect, and conditioning variable.

In our claims data experiments, we performed all conditional Granger causality tests only on subpopulations of patients in whom the potential Granger effect occurred. Thus, we required a Granger cause to predict the timing of a Granger effect in patients who had the Granger effect at some time. This requirement implicitly adjusts conditional Granger causality tests for *baseline* confounders that impact the probability of occurrence of the Granger cause and Granger effect but do not impact their relative timing.[1]

[1]This adjustment was necessary because when we applied our test to the full population, conditional Granger non-causality was extremely rare. Every health event was temporally associated with every other health event conditional on at most one additional health event. The reason appeared to be that some patients generally had many more conditions of all types than other patients, i.e. some patients were frailer at baseline than others. This led to associations between all conditions because occurrence of any given condition was more likely in a frail patient, and such patients in turn were more likely to have other conditions in their records. Restricting Granger causality tests to subpopulations who at some time had the potential Granger effect made conditional Granger non-causality much more common.

Any causal conclusion from $\mathcal{O}$ rests on multiple conditional Granger causality tests. To conclude that $\mathcal{O}$ implies that O_2 causes O_3, recall that the following conditional Granger causality relations must hold:

$$O_1 \rightarrow_{GC} O_2|\emptyset \tag{2.14}$$

$$O_1 \rightarrow_{GC} O_2|O_3 \tag{2.15}$$

$$O_2 \rightarrow_{GC} O_3|\emptyset \tag{2.16}$$

$$O_2 \rightarrow_{GC} O_3|O_1 \tag{2.17}$$

$$O_1 \rightarrow_{GC} O_3|\emptyset \tag{2.18}$$

$$O_1 \not\rightarrow_{GC} O_3|O_2 \tag{2.19}$$

And to say that $\mathcal{O}$ implies that O_2 is just spuriously associated with O_3, recall that in addition to (14)-(17) above, the following conditional Granger non-causality relation must hold:

$$O_1 \not\rightarrow_{GC} O_3|\emptyset \tag{2.20}$$

In our experiments, to be conservative in avoiding causal errors, we set $\delta = .001$. Thus we required that the posterior probability of any model favoring Granger causality be greater than .999 in order to declare Granger causality or less than .001 in order to declare Granger non-causality.

2.3.3 Step 3: Tallying the Results

The output of EGG comprises the causal implications of 50 subprocesses. Most subprocesses do not resolve the association of interest. Say Q subprocesses are resolving, T point to a real association, and F point to a spurious association. A convenient summary statistic of the strength of evidence about the causal nature of the association between events $\mathbf{e}_i$ and $\mathbf{e}_j$ is P* as defined below:

$$\begin{aligned} &\text{If } \mathbf{e}_i \rightarrow_{GC} \mathbf{e}_j|\emptyset \text{ and } \mathbf{e}_i \rightarrow_{GC} \mathbf{e}_j|\mathbf{e}_h, P^*(\mathbf{e}_i, \mathbf{e}_j) = \frac{T+1}{T+F+2} \\ &\text{Otherwise, } P^* = 0. \end{aligned} \tag{2.21}$$

If $Q > 0$ and $T \sim Binomial(Q, p)$, P* would be the posterior mean of p assuming a $Beta(1, 1)$ prior. Higher values of P*($\mathbf{e}_i$,$\mathbf{e}_j$) tend to support causality, and lower values tend to support spurious association. P* is not the posterior probability of causation, however. It is not even reasonable to assume that $T \sim Binomial(Q, p)$ because the conclusions of various subprocesses are not necessarily independent. P* is just a convenient way to compare the strength of evidence from EGG outputs for different event pairs that have different numbers of resolving subprocesses.

Because the distribution of P* is impossible to reliably determine theoretically due to inevitable model misspecification, we recommend constructing empirical distributions of true negative controls and true positives for testing (Schuemie et al, 2013). For example, to evaluate whether stroke causes paralysis, one could first find a collection of true negative conditions $\{\mathbf{e}_1^{TN}, \ldots, \mathbf{e}_V^{TN}\}$ that stroke does not cause but does tend to precede and apply EGG to each of these to obtain $\{P^*(stroke, \mathbf{e}_1^{TN}), \ldots, P^*(stroke, \mathbf{e}_V^{TN})\}$. One could similarly collect a a group of

true positive conditions that stroke does cause, apply EGG to these, and obtain $\{P^*(stroke, \mathbf{e}_1^{TP}), \ldots, P^*(stroke, \mathbf{e}_{V'}^{TP})\}$. Finally, compare $P^*(stroke, paralysis)$ to the distributions of true negatives and true positives to assess the probability that it comes from the true positive distribution.

2.3.4 A Real Example: Ischemic Stroke and Paralysis

We collected 50 precursors of stroke meeting the criteria laid out in Section 3.1. For each precursor $\mathbf{e}_h$, we performed a series of conditional Granger causality tests on the subprocess $\{\mathbf{e}_h, stroke, paralysis\}$ as described in Section 3.2. 9 out of the 50 subprocesses resolved the association between stroke and paralysis, and 8 of the 9 implied causality (P*=.82). That is, there were 8 precursors of stroke that were temporally associated with paralysis but almost entirely through stroke, providing strong evidence that stroke causes paralysis. We then computed P*(stroke,X) for multiple conditions X that tend to follow stroke and are either known to be caused by stroke or known not to be caused by stroke. P*(stroke,paralysis) was firmly in the distribution of P* values for the true positives and separated from the distribution of P* values for the true negatives.

2.4 A Simulation

In this section, we describe the results of applying EGG to simulated longitudinal data with unobserved confounding. The simulation setup was meant to resemble the structure of a health history, though of course it was vastly less complex. We generated 10,000 simulated patient records, each observed for 50 'days' and monitored for 100 'conditions'. Each condition was caused by two other randomly

selected conditions and was only allowed to occur once. (The restriction to a single occurrence was meant to mimic the common practice when working with claims data of only counting the first occurrences of events. This practice is in response to the fact that subsequent appearances of an event in a patient record are likely to be billing artifacts that do not actually correspond to new occurrences.) A diagram of the causal connections between conditions is in Figure 2 below. When a cause occurs for the first time, it increases the probability of its effect occurring by a constant amount for each day afterwards, until the effect occurs at which point its probability of occurring again drops to 0. The probability of event j occurring in patient p at time t was given by

$$P(e_{pj}(t) = 1|H_p^{t-1}) = logit^{-1}(-4 + 2max(c_{pj1}^{t-1}) + 2max(c_{pj2}^{t-1})), \text{ if } max(e_{pj}^{t-1}) = 0; \\ 0, \text{ otherwise} \tag{2.22}$$

where c_{pji}^{t-1} represents the history of the i^{th} cause of condition j in patient p through time $t-1$.

We identified a set of conditions that were linked by common direct causes, and therefore fairly strongly associated, but were not connected to each other by any directed path in the causal diagram. The associations between these conditions were purely spurious. We then dropped the common causes from the dataset so that we were left with 28 condition pairs associated due to unobserved confounding. Since the temporal association went in both directions for these confounded pairs, we arbitrarily assigned one condition in each pair to be a potential cause and

the other to be a potential effect. We then applied EGG to these 28 true negative pairs and to 50 additional true positive pairs that were directly causally linked.

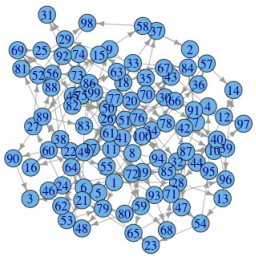

Figure 2.2: A causal graph of the 100 'conditions' in our simulated health records

2.4.1 Selecting Subprocesses

We screened for candidate precursors using logistic regression. For each potential cause e_c, for each potential precursor e_b (i.e. every condition other than e_c), we fit the logistic regression

$$e_{pc}(t) \sim max(e_{pb}^{t-1}) + max(e_{pc}^{t-1}) \tag{2.23}$$

For each pair $\{e_c, e_d\}$, where e_d is the potential effect, we then collected the 10 potential precursors $\{prec_{c1}, \ldots, prec_{c10}\}$ with p-values less than .01 that had largest logistic regression coefficients predicting e_c and formed subprocesses $\{\mathcal{O}_1, \ldots, \mathcal{O}_{10}\} = \{\{prec_{c1}, e_c, e_d\}, \ldots, \{prec_{c10}, e_c, e_d\}\}$.

2.4.2 Conditional Granger Causality Testing

We performed a modified version of the conditional Granger causality tests described in section 3.2 without restricting to the subpopulation in which the potential Granger effect occurs. It was necessary to modify the test because our process was not stationary. (The value of each variable at time t depends on the entire histories of its two causes, and the length of those histories varies with t.) Stationarity is a technical requirement for the validity of Eichler's framework. Without stationarity, the history of a variable that is not at all linked to another variable in the causal network (i.e not linked through a directed path or through a common ancestral cause) can nonetheless carry predictive information by acting as a rough clock indicating the likely value of t. To address this problem, in each PCIM we included an additional binary predictor process indicating whether $t > 15$. We chose 15 because that was near the mode of the distribution of first event times. Conditioning on this additional variable prevented the conditional Granger causality tests from picking up extraneous dependencies unrelated to the causal network.

2.4.3 Results

There was striking separation between the values of P* for the true negatives and the true positives. While the procedure occasionally drew incorrect conclusions from individual subprocesses, the ensemble was never badly fooled by unobserved confounding and had good power to detect real causal effects relative to the empirical null distribution. Note that any standard method assuming ignorability would identify every true negative (along with every true positive) as a causal relation-

ship in this setup. Also note that this simulation demonstrates some robustness to violations of stationarity of the full process. The R code for these simulations is available at zshahn.columbia.edu.

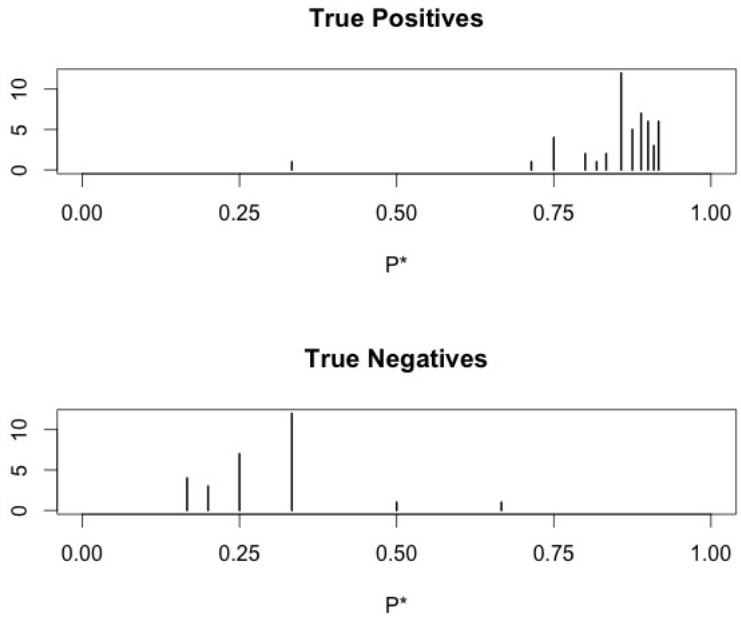

Figure 2.3: Output from applying EGG to 50 true causal effects and 28 spurious associations in simulated data

2.5 Experiments With Claims Data

We explored the performance of EGG on several problems in a real medical claims database. We evaluated EGG's ability to separate (a) effects from non-effects of ischemic stroke; (b) causes from non-causes of ischemic stroke; and (c) drugs for which acute renal failure is a side effect from drugs for which acute renal failure is not a side effect. Using the empirical distributions generated in (b), we applied

EGG to two questions of genuine interest in stroke research: (1) Does sepsis cause ischemic stroke?; and (2) Does sleep apnea cause ischemic stroke? We compared the performance of EGG on all of these problems to a high dimensional inverse propensity weighted cohort method.

All applications were performed in the Thomson Reuters MarketScan Lab Database (MSLR) which contained approximately 8 million patients. The database was converted to the OHDSI Common Data Model (CDM) (Stang et al., 2010). We defined conditions by mapping their ICD-9 codes to the corresponding concept IDs in the CDM. Unfortunately, it was beyond the scope of this project to implement more complex and better validated definitions for each condition. We included only the first occurrence of any condition in a health record. The reason is that repeat occurrences tend to be artifacts of billing practices that do not represent actual reoccurrences of the health event.

2.5.1 Effects of Ischemic Stroke

We compiled a list of conditions that were strongly predicted by ischemic stroke (incidence ratio > 2) in univariate analyses and occurred at least 25,000 times in our data. Working with a stroke neurologist, we identified a subset of these stroke successors that are definitely caused by stroke and a subset that are definitely not caused by stroke. The incidence ratios of the true positives were mostly higher than the incidence ratios of the true negatives. For each true positive and true

negative condition we identified, we applied EGG to test whether it was caused by stroke, proceeding as described in Section 3. Examples of stroke precursors that our algorithm selected are brain neoplasm, cerebral ischemia, carotid artery obstruction, dementia, Alzheimer's disease, Parkinson's disease, heart failure, kidney failure, hemiplegia, and atrial fibrillation. The separation of P* between true positives and true negatives was striking, similar to the results of the simulation study. The full output including numbers of resolving subprocesses implying causality and spurious association is in the appendix.

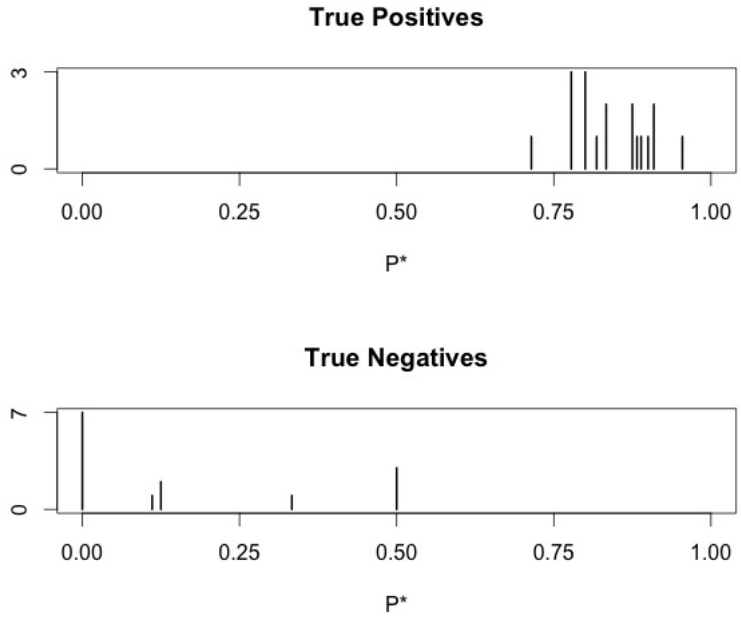

Figure 2.4: Output from applying EGG to 18 true effects and 14 spurious successors of ischemic stroke

2.5.2 Causes of Ischemic Stroke

We compiled a list of conditions that strongly predicted ischemic stroke (incidence ratio > 2) in univariate analyses in our data. Working with a stroke neurologist,

we identified a subset of these stroke precursors that definitely cause stroke and a subset that definitely do not cause stroke. The distribution of incidence ratios was similar in the true positive and true negative groups, with the lowest few incidence ratios actually being true positives. For each true positive and true negative condition we identified, we applied EGG to test whether it causes stroke, proceeding as described in Section 3. The separation is not as good for this application as for the effects of stroke. This may be because causal effects are weaker and responsible for a smaller proportion of outcome occurrences. Still, EGG produced stronger evidence of causality for 6 of the 13 true positives than any of the 19 true negatives. This was far superior to the comparator cohort method, as we will see later.

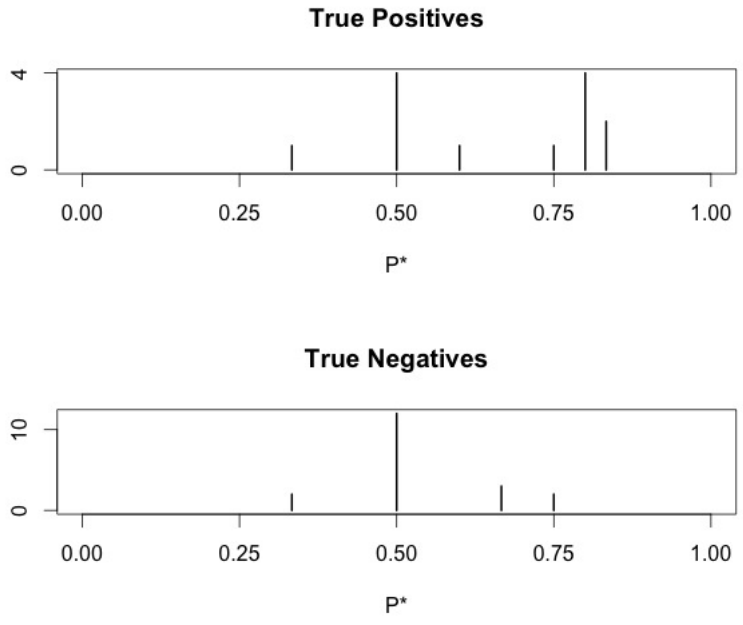

Figure 2.5: Output from applying EGG to 13 true causes and 19 spurious precursors of ischemic stroke

2.5.3 Two Questions of Interest

Strong temporal associations between both sepsis and sleep apnea and ischemic stroke have been observed (Walkey et al., 2011; Yaggi et al., 2005). There are plausible mechanisms for a causal connection in both cases, but causality is not firmly established. Using the empirical true negative and true positive distributions generated in Section 5.3 by the stroke precursors, we applied EGG to investigate whether sepsis or sleep apnea cause ischemic stroke.

For sepsis, 8 of 50 subprocesses resolved the association, and all 8 implied a causal relationship (P*=.9). This was much stronger evidence for causality than any of the 19 true negatives (or true positives). These results say that there are many conditions that tend to precede sepsis and are only associated with ischemic stroke through sepsis. This is strong evidence for causality.

But taking a look at the CDM concept names of the 8 resolving precursors may reduce the strength of our conclusion. Those concept names are: 'Osteomyelitis of ankle AND/OR foot', 'Gangrenous disorder', 'Infection AND/OR inflammatory reaction', 'ulcer of heel', 'ulcer of knee', 'Esophageal varices without bleeding', 'Osteomyelitis', and 'Ulcer of lower limb'. While these are all technically distinct concepts in our data, it seems silly to count some of them as separate precursors offering independent sources of evidence. For example, 'Osteomyelitis' and 'Osteomyelitis of ankle AND/OR foot' are surely very similar. As are 'Ulcer or lower limb', 'Ulcer of heel', and 'Ulcer of knee'. But even if sepsis only has 4 or 5 precursors identifying it as causal (and none identifying it as spurious), this is

solid evidence for causality in light of the empirical distributions from the other stroke precursors.

For sleep apnea, there were only 25 precursors that met our criteria for strength of association, prevalence, and distinctness from the potential cause. Of these 25, three resolved the association between sleep apnea and stroke, and all three implied a causal relationship (P*=.8). This is stronger evidence of causality than EGG produced for any of the negative controls. The three resolving precursors were narcolepsy, movement disorder (e.g. the sleeping disorder restless leg syndrome), and sarcoidosis.

The full output for sepsis and sleep apnea are in the appendix.

2.5.4 A Comparator Cohort Method

We also performed cohort analyses for all of the stroke examples. For each analysis, we randomly selected a control cohort of the same size as the exposed cohort, taking the index date for the control cohort to be a random hospital visit. We then used lasso logistic regression to fit high dimensional propensity score models predicting cohort membership based on binary predictors indicating pre-index date occurrence of every condition, drug, and procedure as well as age and gender and a few frailty measures (total pre-index date conditions, drugs, and hospital visits). We used the estimated propensity scores to fit inverse propensity score weighted cox survival models for the outcome with cohort as the predictor. For

stability, as is common practice, we trimmed the data to only include observations with propensity scores between .05 and .95.

The cohort method had very poor positive predictive value as typically applied. With p-value less than .01 as the criterion for causality, the cohort comparator detected a causal effect for every condition that definitely did not cause stroke and 13 of the 14 conditions that stroke definitely did not cause. (Recall that all the negative controls were chosen to be very strongly associated with stroke in unadjusted analyses.)

We also looked at whether there was separation in the empirical distributions of estimated adjusted hazard ratios between true positives and true negatives. There was no separation between causes and non-causes of stroke.

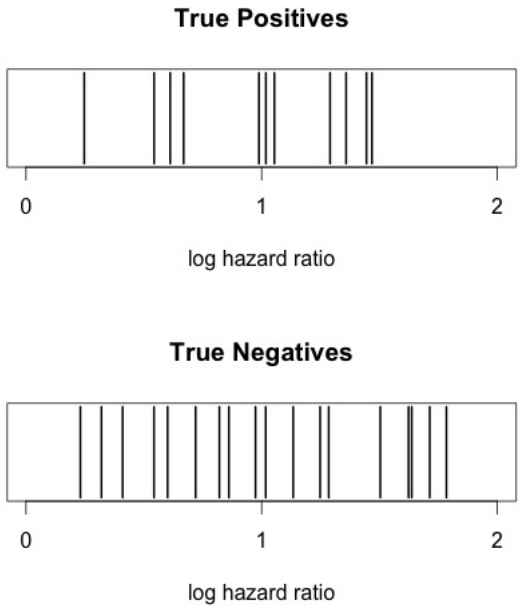

Figure 2.6: Adjusted log hazard ratios of 13 true causes and 19 spurious precursors of ischemic stroke

There was much better separation for effects and non-effects of stroke, but there was also some separation in the unadjusted incidence ratios. The cohort method did not distinguish well between the true positives and true negatives with similar unadjusted incidence ratios, leading to some overlap between the two distributions where EGG had none (see Figure 4).

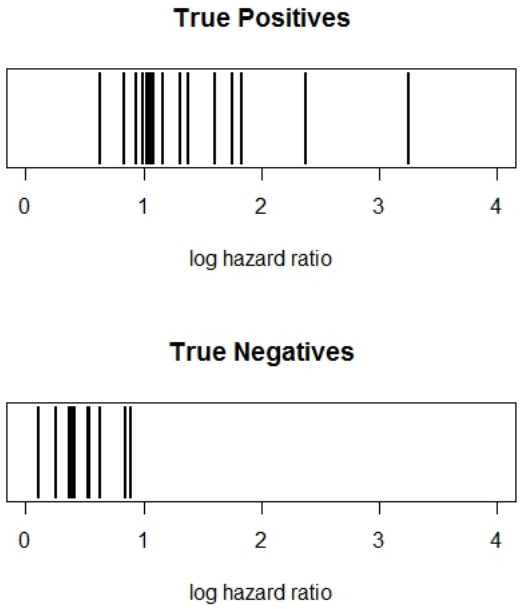

Figure 2.7: Adjusted log hazard ratios of 18 true effects and 14 spurious successors of ischemic stroke

2.5.5 Drug Side Effects

In a series of empirical experiments (called the OMOP experiments), a group of researchers recently evaluated the performance of many standard epidemiological methods at detecting side effects of drugs using claims data (Madigan et al., 2014). They found that confidence intervals for all methods had poor coverage probability on a set of negative controls where the true effect size was thought to be zero. Such poor coverage leads to high false positive rates and low positive predictive value for tests based on individual analyses. When tests were based on empirical distributions of positive and negative controls, results were mixed. For a given side effect and database, there usually existed settings for some method that separated the positive and negative controls very well. However, slightly different settings

of the best separating method often led to poor separation in the same database. And methods that separated positives and negatives well for one combination of database and side effect often performed very poorly for other combinations.

Among the positive and negative controls considered in the OMOP experiments were 88 drugs that were either known to cause or not to cause acute renal failure. 24 of the 88 drugs were true positives. The best performing method in the MSLR database (an earlier version of the same database we used in this study) was a self controlled case series approach that achieved perfect separation with an AUC of 1. But performance was sensitive to method settings such as exposure window as well as method type, and many standard methods had AUCs no better than random guessing.

We applied EGG to this set of 88 drugs. EGG did not provide evidence of causality for any drug or separate the true positives from the true negatives. EGG's poor power in this setting is presumably due to the fact that the true causal effects are weak. Serious drug side effects tend to be very rare; otherwise, they would be detected in clinical trials and the drug would not be allowed on the market. EGG searches for associations between precursors and effects that are only transmitted *through* causes. When causal effects are weak, indirect associations that are only transmitted through the causal effects are even weaker and therefore very difficult for our crude conditional Granger causality tests to detect. There were only 84,626 instances of acute renal failure in the (relatively small) database we were using for our experiments, and it is possible that EGG would perform better at detecting weak effects in a larger database.

2.6 Discussion

The key idea behind this paper is that if it is necessary to condition on a potential cause to block the temporal association between its precursor and a potential effect, this can be construed as evidence of causality. Also, if a potential cause's precursor is unassociated with a potential effect unconditional on the potential cause, this can be construed as evidence for spurious association. This reasoning (let us call it 'blockage reasoning') is formally justified under certain assumptions by Michael Eichler's framework of Granger causal graphical models and dMAGs, but it is also intuitive. True causal effects ought to 'pass forward' temporal associations in time, whereas spurious associations ought not to. Whether it is necessary to condition on a potential cause in order to block the association between its precursor and its potential effect is an indication of whether the potential cause is 'passing forward' the association. Crucially, blockage reasoning cannot be derailed by unobserved confounding alone, unlike standard epidemiological methods for causal discovery. Hence, blockage reasoning can usefully supplement standard techniques by providing an *independent evidence source* susceptible to different biases (Zubizarreta et al., 2012).

We have developed a novel causal discovery algorithm called EGG based on dynamic Maximal Ancestral Graphs that represent conditional Granger (non-)causality relations. To guard against the instability of causal conclusions from individual subprocesses (stemming from possible sensitivity to model misspecification, poor data quality, sampling variation, or departures from technical assumptions about the underlying data generating process), we employ ensembles. Other authors

have applied confounding resistant causal discovery methods to longitudinal data (Entner and Hoyer, 2010; Waldorp et al., 2011; Chu and Glymour, 2008; Eichler, 2007), but we are unaware of any prior work that searches for sets of variables (or ensembles of such sets) to resolve specific associations of interest.

We provided some empirical evidence of EGG's utility from a simulation and claims data experiments in which unobserved confounding was certainly present. Based on these initial findings, EGG appears to be robust to both unobserved confounding and various violations of technical assumptions. Our simulation and experiments both involved non-stationary processes, and the patients in our experiments were heterogeneous at baseline. Yet we saw that EGG rarely produced decisive false positives and had high power to detect strong causal effects. Unfortunately, its power does appear to diminish greatly with effect size. Of course, much more empirical and theoretical work is needed to thoroughly explore the operating characteristics of EGG or other methods built on similar foundations.

We also applied EGG to two questions of interest in stroke research–namely, whether two 'risk factors' for stroke (sleep apnea and sepsis) are actual causes of stroke. In medicine, the term 'risk factor' is often used for any predictor of a condition, whether or not it is causal. It can be important to know whether a risk factor is causal because treating non-causal risk factors serves no preventive purpose. For both sleep apnea and sepsis, we compared EGG's output to the empirical distributions of its output for sets of true positive and true negative causes of stroke. There was very strong evidence that sepsis causes stroke and weaker but still solid evidence that sleep apnea causes stroke.

We take 'EGG' to refer to the general three step process described at the beginning of Section 3. In Step 1, we identify promising subprocesses containing the variables of interest. In Step 2, we employ conditional Granger causality tests to learn the dMAGs representing the conditional Granger causality relations among the subprocesses generated in Step 1 and determine their causal implications (if any). In Step 3, we tally the results of the resolving subprocesses and summarize the evidence. There are many ways one might choose to perform each of these steps and many opportunities to improve over the existing algorithm.

For instance, in Step 3, we count each resolving subprocess separately and equally. But, as we saw in the case of sepsis and stroke, subprocesses may be related to each other. If this is the case, it is inappropriate to consider them as independent sources of evidence. We informally discounted the strength of evidence in our analysis of sepsis, but it should be possible to address the problem more formally. For example, one could develop a weighting scheme based on the distance between precursors in the CDM concept hierarchy (Stang et al., 2010) and/or the empirical associations between precursors in the data.

We also hope to explore alternative conditional Granger causality tests (e.g. tests based on mutual information such as in Amblard and Olivier, 2011). In our simulations and experiments, we made ad hoc modifications to the conditional Granger causality tests in Step 2. Specifically, in our simulations we adjusted for time to deal with lack of stationarity, and in our experiments on claims data we fit models only within subpopulations who experienced the potential Granger effect at

some point in time to handle baseline confounding that impacted occurrence but not relative timing of events. Other ad hoc modifications than those we made may have superior operating characteristics, and it may also be possible to extend Eichler's framework to put EGG on stronger theoretical footing when the sorts of violations and irregularities our modifications were aimed at are present. Also, fast and flexible temporal dependency models for binary time series together with fast and reliable model selection could lead to improved conditional Granger causality tests. Improved tests could potentially allow for the use of larger dMAGs, which could have a higher resolving rate than the 3 variable subprocesses we used and hence increase power.

The most obvious shortcoming of EGG compared to standard methods is that EGG does not estimate the size of causal effects. One could argue that causal discovery is a misguided endeavor as in reality almost everything is causally related to almost everything else, even if the vast majority of causal effects are tiny. In this view, we learn little when we learn that a causal relationship merely exists, the real question being, 'what is its size?'. In practice, however, EGG seems to pick up mainly causal effects that are relatively large and therefore interesting, at least with the sample sizes and thresholds we were working with.

A related concern is that in reality conditional Granger non-causality is exceedingly rare no matter how large the conditioning set. With enough data, we could always detect dependence. If this is so, then for real problems EGG would be anti-consistent in the sense that as sample size increases eventually no subprocess in the ensemble would ever be resolving. However, this concern applies equally

to every existing causal discovery method based on conditional independence constraints.

Another objection one might raise to this work is that our comparison to a cohort method was unfair. The comparator cohort method that relied on ignorability for validity fared considerably worse in our experiments, as it appeared to be consistently foiled by unobserved confounding. We grant that an epidemiologist carefully creating a custom study design and performing a sensitivity analysis for each example would have likely performed better than our cohort method. We also note that EGG was not carefully applied, either. In the hands of an epidemiologist, careful selection of precursors based on subject matter knowledge might improve performance.

The main contribution of this paper was to demonstrate that an algorithm based on blockage reasoning could be successfully applied to high dimensional longitudinal data for causal discovery and avoid some of the pitfalls related to unobserved confounding that plague more standard methods. We made many specific choices in constructing our particular algorithm, the use of ensembles probably being the most important. But there are many other ways one could have proceeded. We hope these preliminary results encourage exploration of similar (and better) approaches.

2.7 Bibliography

Amblard, Pierre-Olivier, and Olivier JJ Michel. "On directed information theory and Granger causality graphs." Journal of computational neuroscience 30.1 (2011): 7-16.

Chu, Tanjiao.and Glymour, Clark. "Search for Additive Nonlinear Time Series Causal Models," Journal of Machine Learning Research 9 (2008) 967-991

Eichler, Michael, and Vanessa Didelez. "On Granger causality and the effect of interventions in time series." Lifetime data analysis 16.1 (2010): 3-32.

Eichler, Michael. "A graphical approach for evaluating effective connectivity in neural systems." Philosophical Transactions of the Royal Society of London B: Biological Sciences 360.1457 (2005): 953-967.

Eichler, Michael. "Granger causality and path diagrams for multivariate time series." Journal of Econometrics 137.2 (2007): 334-353.

Eichler, Michael. "Graphical modelling of multivariate time series." Probability Theory and Related Fields 153.1-2 (2012): 233-268.

Eichler, Michael. "Causal inference from time series: What can be learned from granger causality?." Proceedings of the 13th International Congress of Logic, Methodology and Philosophy of Science. 2007.

Eichler, Michael. "Graphical Gaussian modelling of multivariate time series with latent variables." International Conference on Artificial Intelligence and Statistics. 2010.

Eichler, Michael. "Causal inference in time series analysis." In: C. Berzuini, A.P. Dawid, L. Bernardinelli (eds), Causality: Statistical Perspectives and Applications, Wiley, Chichester. 2012

Eichler, Michael. "Causal inference with multiple time series: principles and problems." Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 371.1997 (2013): 20110613.

Entner, Doris, and Patrik O. Hoyer. "On causal discovery from time series data using FCI." Probabilistic graphical models (2010): 121-128.

Gunawardana, Asela, Christopher Meek, and Puyang Xu. "A model for temporal dependencies in event streams." Advances in Neural Information Processing Systems. 2011.

Harrold LR, Saag KG, Yood RA, Mikuls TR, Andrade SE, et al. 2007. Validity of gout diagnoses in administrative data. Arthritis Rheum. 57:103?8

Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. 2007. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol. Drug Saf. 16:393?401

Madigan, David, et al. "A systematic statistical approach to evaluating evidence from observational studies." Annual Review of Statistics and Its Application 1 (2014): 11-39.

Meek, Christopher. "Toward Learning Graphical and Causal Process Models." UAI 2014 Workshop Causal Inference: Learning and Prediction. 2014.

Pearl, J., 2009. Causality. Cambridge University Press, Cambridge, UK.

Richardson, Thomas S., and James M. Robins. "Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128 (2013).
Richardson, Thomas, and Peter Spirtes. "Ancestral graph Markov models." Annals of Statistics (2002): 962-1030.

Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.

P. E. Stang, P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock, Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership, Ann Intern Med 153(9) (2010), 600 ? 606.

Waldorp, Lourens, Ingrid Christoffels, and Vincent van de Ven. "Effective connectivity of fMRI data using ancestral graph theory: dealing with missing regions." NeuroImage 54.4 (2011): 2695-2705. (Uses ancestral graphs to avoid confounding, but doesn't take time into account)

Walkey AJ, Wiener RS, Ghobrial JM, Curtis LH, Benjamin EJ. Incident stroke and mortality associated with new-onset atrial fibrillation in patients hospitalized with severe sepsis. JAMA. 2011;306:2248-2254.

White, Halbert, and Xun Lu. "Granger causality and dynamic structural systems." Journal of Financial Econometrics 8.2 (2010): 193-243.

Wild, Beate, et al. "A graphical vector autoregressive modelling approach to the analysis of electronic diary data." BMC medical research methodology 10.1 (2010): 28.

Yaggi, H. Klar, et al. "Obstructive sleep apnea as a risk factor for stroke and death." New England Journal of Medicine 353.19 (2005): 2034-2041.

Zubizarreta, Jose R., et al. "Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia." Journal of the American Statistical Association 107.499 (2012): 901-915.

2.8 Appendix A: Effects of Stroke

Conditions	$stroke \rightarrow_{GC} \mathbf{e}\|\emptyset$	#GnC	spurious	causal	P*	truth
altered mental	1.000	0	0	8	0.004	1
speech	1.000	0	1	15	0.000	1
lack of coordination	1.000	0	1	14	0.000	1
gait	1.000	0	0	9	0.002	1
resp failure	1.000	0	0	3	0.125	1
psychotic	1.000	0	1	6	0.062	1
septicemia	1.000	0	0	4	0.062	1
strength reduced	1.000	0	0	6	0.016	1
amnesia	1.000	0	0	9	0.002	1
venous thrombosis	1.000	0	1	4	0.188	1
ami	1.000	0	0	3	0.125	1
hip injury	1.000	0	0	3	0.125	1
pulmonary embolism	1.000	0	1	6	0.062	1
involuntary movement	1.000	0	1	6	0.062	1
bedsores	1.000	0	0	6	0.016	1
cerebral edema	1.000	0	0	20	0.000	1
paralysis	1.000	0	1	8	0.020	1
pneumonitis	1.000	0	1	9	0.011	1
cataract	0.003				1.000	0
glaucoma	1.000	14	7	0	0.992	0
spinal stenosis	0.814					0
acute renal failure	0.964	25	10	0	0.999	0
acidosis	0.000				1.000	0
kidney stage3	1.000	0	6	0	0.984	0
abdominal aneurism	0.000				1.000	0
emphysema	1.000	1	6	0	0.984	0
foot ulcer	0.000				1.000	0
diabetic polyneuropathy	0.006				1.000	0
amd	0.001				1.000	0
diabetic renal	0.000				1.000	0
hypertension	1.000	0	1	0	0.500	0
diabetic oculopathy	0.742					0

2.9 Appendix B: EGG Output for Sleep Apnea

Precursor	IR	Prevalence	$\mathbf{e}_h \rightarrow_{GC} stroke \vert \emptyset$	$apnea \rightarrow_{GC} stroke \vert \mathbf{e}_h$	$\mathbf{e}_h \rightarrow_{GC} stroke \vert apnea$	conclusion
Narcolepsy	3.553	0.001	0.994	1.000	0.000	1
Movement disorder	2.799	0.004	1.000	1.000	0.000	1
Morbid obesity	2.776	0.035	1.000	1.000	1.000	
PPHT	2.565	0.002	1.000	1.000	1.000	
pulmonary heart disease	2.485	0.005	1.000	1.000	1.000	
Articular gout	2.392	0.004	0.123	1.000	0.000	
Respiratory observation	2.366	0.078	1.000	1.000	1.000	
Hypoxaemia	2.331	0.010	1.000	1.000	1.000	
Congestive cardiac failure	2.305	0.018	1.000	1.000	1.000	
Acute asthmatic bronchitis	2.242	0.002	1.000	1.000	1.000	
Sarcoidosis	2.151	0.002	1.000	1.000	0.000	1
Cardiomyopathy	2.135	0.010	1.000	1.000	1.000	
coronary artery bypass graft	2.135	0.002	1.000	1.000	0.046	
Left ventricular failure	2.134	0.003	1.000	1.000	1.000	
Vaquez's disease	2.127	0.002	1.000	1.000	1.000	
Atrial fibrillation	2.121	0.019	1.000	1.000	1.000	
cardiac defibrillator in situ	2.091	0.002	1.000	1.000	1.000	
Atrial flutter	2.085	0.004	1.000	1.000	1.000	
Hypertensive CHF	2.072	0.001	1.000	1.000	1.000	
Abnormal cardiovascular	2.044	0.008	1.000	1.000	1.000	
Diastolic heart failure	2.042	0.002	1.000	1.000	1.000	
Neurologic disorder diabetes	2.032	0.014	1.000	1.000	1.000	
Coronary arteriosclerosis	2.031	0.044	1.000	1.000	1.000	
Cardiomegaly	2.030	0.020	1.000	1.000	1.000	
Preinfarction syndrome	2.024	0.012	1.000	1.000	1.000	

2.10 Appendix C: EGG Output for Sepsis

Precursor	IR	Prevalence	$\mathbf{e}_h \rightarrow_{GC} stroke \vert \emptyset$	$sepsis \rightarrow_{GC} stroke \vert \mathbf{e}_h$	$\mathbf{e}_h \rightarrow_{GC} stroke \vert sepsis$	conclusion
kidney stage 5	15.812	0.001	1.000	1.000	1.000	
Dialysis observation	15.458	0.001	1.000	1.000	1.000	
Complication of implant	14.835	0.002	1.000	1.000	1.000	
End stage renal disease	14.727	0.003	1.000	1.000	1.000	
Erythroid aplasia	14.142	0.001	1.000	1.000	0.183	
Anemia in neoplastic disease	13.594	0.001	1.000	1.000	1.000	
Pancytopenia	13.406	0.002	1.000	1.000	1.000	
hyperparathyroidism of renal origin	13.053	0.002	1.000	1.000	1.000	
Anemia of chronic renal failure	12.933	0.004	1.000	1.000	1.000	
Pressure sore	12.547	0.001	1.000	1.000	0.984	
Liver secondary cancer	12.492	0.002	1.000	1.000	1.000	
Alcoholic cirrhosis	12.082	0.001	1.000	1.000	1.000	
Portal hypertension	11.945	0.001	1.000	1.000	0.506	
Osteomyelitis of ankle AND/OR foot	11.770	0.001	1.000	1.000	0.000	1
Acute tubular necrosis	11.677	0.002	1.000	1.000	1.000	
Chronic kidney disease stage 4	11.348	0.003	1.000	1.000	1.000	
Gangrenous disorder	11.288	0.001	1.000	1.000	0.000	1
Deficiency of macronutrients	11.145	0.005	1.000	1.000	1.000	
Secondary malignant tumour of lung	11.082	0.002	1.000	1.000	1.000	
Infection from device, implant AND/OR graft	10.998	0.003	1.000	1.000	0.000	1
Acquired thrombocytopenia	10.841	0.002	1.000	1.000	1.000	
ARF - Acute renal failure	10.841	0.012	1.000	1.000	1.000	
Acute-on-chronic respiratory failure	10.812	0.001	1.000	1.000	1.000	
Metastatic tumour of bone	10.758	0.002	1.000	1.000	1.000	
Ulcer of heel	10.696	0.002	1.000	1.000	0.000	1
Chronic respiratory failure	10.633	0.001	1.000	1.000	1.000	
Malignant neoplasm of liver	10.502	0.001	1.000	1.000	1.000	
Acute on chronic systolic heart failure	10.471	0.001	1.000	1.000	1.000	
secondary brain cancer	10.441	0.001	1.000	1.000	1.000	
Pneumonitis due to inhalation of food	10.416	0.002	1.000	1.000	1.000	
Ulcer of knee	10.392	0.001	1.000	1.000	0.000	1
RF - Renal failure	10.342	0.005	1.000	1.000	1.000	
Acute respiratory failure	10.295	0.008	1.000	1.000	1.000	
Secondary malignant neoplastic disease	10.265	0.002	1.000	1.000	1.000	
Esophageal varices without bleeding	10.140	0.001	1.000	1.000	0.001	1
Obstruction of bile duct	10.090	0.001	1.000	1.000	1.000	
Osteomyelitis	10.079	0.002	1.000	1.000	0.000	1
postoperative pulmonary insufficiency	10.058	0.002	1.000	1.000	1.000	
Ulcer of lower limb	9.949	0.003	1.000	1.000	0.000	1
Metabolic encephalopathy	9.921	0.001	1.000	1.000	1.000	
lung cancer	9.876	0.001	1.000	1.000	1.000	
Dementia from another disease	9.849	0.002	1.000	1.000	1.000	
Pulmonary oedema - acute	9.842	0.001	1.000	1.000	1.000	
Drug-induced neutropenia	9.813	0.002	1.000	1.000	1.000	
Hypertensive renal disease with renal failure	9.789	0.010	1.000	1.000	1.000	
PAIN OF METASTATIC MALIGNANCY	9.772	0.001	1.000	1.000	1.000	
throat cancer	9.568	0.003	1.000	1.000	1.000	
Acidosis	9.564	0.005	1.000	1.000	1.000	
Anaemia of chronic disease	9.555	0.004	1.000	1.000	1.000	
Acute diastolic heart failure	9.406	0.001	1.000	1.000	1.000	

Chapter 3

Predicting Health Outcomes from High Dimensional Longitudinal Health Histories Using Relational Random Forests

3.1 Introduction

With increasingly widespread use of Electronic Health Records (EHRs), predicting health outcomes from high dimensional, longitudinal health histories is of central importance to healthcare. The medical literature has formalized such prediction problems in a few instances, and the resulting "risk calculators" attract widespread use.

CHADS2 is one example of a risk calculator that is currently used in practice to predict ischemic stroke in patients with atrial fibrillation (afib). The goal of prediction with CHADS2 is to identify patients with sufficiently low risk of stroke to be spared warfarin, an anticoagulant with well-known side effects that also requires extensive monitoring. Like most widely used risk calculators, the statistical

analysis underlying CHADS2 makes use of a small fraction of the available patient-level information.

We use health insurance claims data from the Observational Medical Outcomes Partnership (OMOP) to construct our own high dimensional risk calculators for stroke in afib patients. (OMOP is a public private partnership that developed methodological research experiments to study the performance of analysis methods in observational health data, both claims data and EHRs [1,2,3].) We first apply standard machine learning approaches such as L1 regularized logistic regression [4] and random forests [5,6] to naive mappings of patient histories into simple (albeit high-dimensional) feature vectors containing coarse temporal information about occurrences of health events. These methods demonstrate superior predictive performance to CHADS2 in our database but fail to identify a sizable population with very low risk of stroke.

We hypothesize that incorporating (potentially complex, high order) temporal relations between health events may further improve predictive performance. The challenge is to identify informative members of the vast set of such relations. To this end, we adapt a predictive modeling method (referred to in the remainder as Relational Random Forests, or RRF) originally developed in the context of speech recognition [7]. RRF greedily constructs informative labeled graphs representing temporal relations between multiple health events at the nodes of randomized decision trees. We modify the existing algorithm so that the pool of candidate temporal relations is determined by the data rather than specified (necessarily arbitrarily, in our case) by the analyst.

Many other researchers have considered the problem of developing predictive models from medical histories. Past efforts may be categorized according to their approach to variable selection. In the vast majority of cases, researchers generated a small set of candidate covariates thought to be related to the outcome of interest and employed traditional model selection algorithms such as stepwise regression to choose among them. The CHADS2 score is a product of this approach, as are other well known risk scores such as those arising from the Framingham Heart Study.

Some researchers, like us, have taken a machine learning approach to variable selection. That is, they define a very large set of features, few of which are a priori likely to be related to the outcome, and employ algorithms such as regularized regression or CART that automatically incorporate informative subsets of the features into predictions. For example, [8] predicts hospital readmissions using logistic regression with a modified forward variable selection algorithm to choose features from the set of all indicator variables for any past occurrence of a medical concept with an ICD-9 code. None of the many features used in [8], however, conveys any information about time.

[9] did recently address the problem of incorporating temporal patterns from medical histories into predictive models. They consider a patient's health history as the superposition and concatenation of multiple pattern matrices. Each pattern matrix specifies a rigid temporal relationship among health events that repeats over time. They employ a matrix factorization algorithm (called OSC-NMF) to learn

the pattern matrices, which can then be used to construct features for predictive models. RRFs are able to incorporate more flexible temporal relationships than can be represented by pattern matrices, and the sample application of OSC-NMF in [9] only involved 30 candidate health events compared to over 10,000 in our application. We are curious about how OSC-NMF scales, and we hope to explore this exciting approach more deeply in future work.

It is rare for a machine learning feature set to include temporal relationships between large numbers of health events. The reason is that there are too many such relationships for standard machine learning algorithms to identify informative subsets. The advantage of RRF is that it allows feature sets to effectively include a wide class of temporal relationships.

3.2 Methods

Our data comprised detailed time-stamped insurance claims records of patient-level health events (e.g. drugs prescribed, conditions diagnosed, procedures performed, etc.) occurring over a span of several years from the MarketScan Multi-State Medicaid database. In patients with atrial fibrillation and at least one year of continuous observation period both preceding and following their first observed afib diagnosis, we predicted the occurrence of stroke in the year following the index afib diagnosis using multiple approaches. We considered patients who died to be fully observed and thus only excluded patients who left the Medicaid program for reasons other than death or who were diagnosed with afib less than a year before the latest observation date in the database. With the above noted restrictions,

CHADS2 Score	Estimated Probability of Stroke
0	.019
1	.028
2	.04
3	.059
4	.085
5	.125
6	.182

Table 3.1: Estimated probability of stroke corresponding to each CHADS2 score

our dataset consisted of 12,581 patients, 1,850 of whom had strokes.

CHADS2 simply assigns point values for each of five conditions present in a health history: one point each for congestive heart failure, hypertension, age > 75, or diabetes, and two points for prior stroke. Each possible point total is associated with a corresponding estimated probability of stroke in the next year. (See Table 1.) CHADS2 was developed and calibrated using 2580 afib patients randomly assigned to receive aspirin (and no warfarin). The included risk factors were chosen because they were all found to be independently associated with stroke [10]. We calculated a CHADS2 score for each patient in our database and evaluated predictive performance. On account of left censoring of variables that inform the CHADS2 score, we expect that CHADS2 would perform better in a real world setting than in our experiments. In particular, our ability to assess prior stroke is limited to the patient observation period in the database. Similarly for congestive heart failure, hypertension, and diabetes, we are relying on the appearance of diagnostic codes in the medical record during the observation period.

To implement L1 regularized logistic regression and classical random forests, we first encoded health histories into high dimensional binary feature vectors carrying coarse temporal information. We created "time splits" spaced on a log scale at 0, 7, 55, 148, and 435 days before index date of afib diagnosis. For each health event and for each time split, we created two binary predictors—one indicating whether the health event era (e.g. the period of time during which a drug was taken) extended before the time split and one indicating whether the health event era extended after the time split. (See Figure 1 for an illustration.) Including all drugs and conditions at all levels of a hierarchical taxonomy of health events [11], our feature vectors contained over 100,000 binary predictors, as well as age and gender.

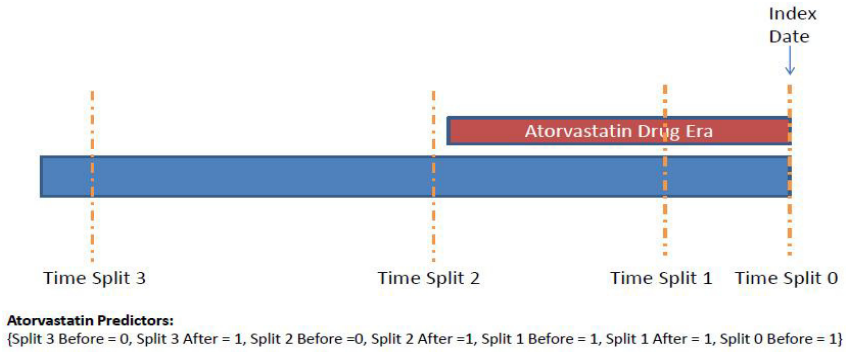

Figure 3.1: An illustration of the encoding of health history information into a set of binary predictors.

We fit standard L1-regularized logistic regression and random forest models to these predictors. We used algorithms that exploited the sparseness of our predictor matrix to cope with its high dimension. For regularized logistic regression we used BBR software [12,13], and for classical random forests we used FEST software [14].

The logistic regression model specifies that

$$\log\left(\frac{P\left(Stroke \mid X\right)}{1-P\left(Stroke \mid X\right)}\right) = \beta' X$$

where X is a vector of predictors representing a patient history and β is a vector of parameters to be estimated. In L1 regularized logistic regression, estimates $\widehat{\beta}$ are obtained as

$$argmin_{\beta} - l\left(\beta\right) + \lambda \left\|\beta\right\|_1$$

where $l(\cdot)$ denotes the log likelihood function and λ is a tuning parameter. The penalty term $\lambda \left\|\beta\right\|_1$ encourages sparse solutions with many coefficients in β set to zero and allows regression models to be fit in scenarios with more predictors than observations. Larger values of λ lead to more sparsity, and its value is chosen through cross validation to minimize an estimate of out of sample prediction error [4].

Random forests, as described in [5], are ensembles of decision trees. Each tree is fit to a random subset of the observations (patients, in our application) and a random subset of predictors. Predictions are calculated as the mean of the predictions of all the trees in the forest. Each tree's prediction for a patient is the proportion of strokes in the terminal node containing that patient. We chose to consider $\sqrt{p}$ potential predictors for each split, where p denotes the total number of predictors, and we grew each tree to a maximum depth of 1000. We grew 1000 trees, at which point predictive performance appeared to have converged.

We hypothesized that incorporating certain complex temporal relationships among multiple health events in the predictive models might lead to more accurate pre-

dictions. The predictors we used for the standard machine learning approaches fail to capture even simple relative temporal relationships of the sort "received drug A then suffered condition B". We can also imagine more specific relationships that may be informative, such as "received drug A then suffered condition B within T days" or "received drug A then suffered condition B all within T days of the index date". Such relationships among multiple variables are well described by labeled graphs with health event types at the vertices and edges labeled to indicate temporal relations between the vertices. Below is an example of a graph representing a set of temporal relationships between health events that occurred in a patient from our data set.

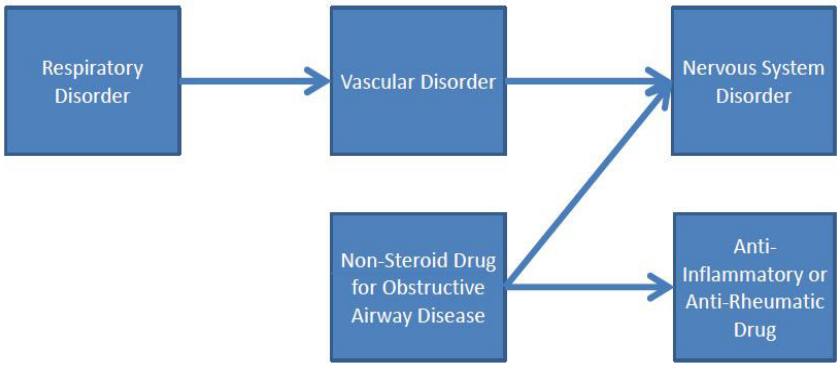

Figure 3.2: A graph representing temporal relations between multiple health events in a health history from our database. Events at the tails of arrows occurred before events at the heads. No information about the time between events or time from events to index date is displayed in this particular graph.

To formalize, let V = $\{e_1,\ldots,e_H\}$ be the set of all possible vertices (that is, the set of all health events that occur in any patient). Then a labeled graph G is of the form $\{(e_{i,1}, e_{i,2}, R_i),\ i = 1,\ldots,|G|,\ e_{i,j} \in V\ for\ j = 1,2\}$ where R_i specifies a

temporal relationship between $e_{i,1}$ and $e_{i,2}$. We call each triplet $(e_{i,1}, e_{i,2}, R_i)$ an edge. We say a health history X *satisfies* an edge $\mathbf{e}_i = (e_{i,1}, e_{i,2}, R_i)$ if $e_{i,1}$ and $e_{i,2}$ occur in X and temporal relation R_i holds between them. We say a health history X satisfies a graph G if it satisfies every edge in G. Let 1_G (respectively, 1_e) denote the random indicator variable that is equal to 1 if a patient's health history satisfies G (respectively, e) and 0 if not.

Our proposed approach grows random decision trees designed to produce informative labeled graphs at the nodes [7]. We call each tree grown in this way a 'relational tree'. Each node of a relational tree is defined by two sets of edges, which we will call I and E for "included" and "excluded". Every patient in a node satisfies each edge in I and none of the edges in E. The edges in I form a labeled graph satisfied by all patients in the node. We grow decision trees by splitting nodes on variables of the form $\mathbf{1}_e$ for $\mathbf{e}$ an edge. Figure 3 depicts a sample tree.

To split a node, we first find a suitable edge $\mathbf{e}$ such that $\mathbf{1}_e$ is predictive of stroke for the node's population. Then we construct two child nodes—one comprising the subpopulation that satisfies $\mathbf{e}$ and the other comprising the subpopulation that does not. Algorithm 1 below describes how to split a node.

Algorithm 1 (Node splitting):
To split a node all of whose patients satisfy every edge in a set I and no edges in a set E:

1. Let S denote the set of candidate edges to split on. Set S = {}

2. While |S|< M for some pre-chosen M (we set M = 40):

   (a) Select a random patient from the node population (we actually enforce that an equal number of cases and non-cases are selected).

   (b) Using algorithm 1A (below), select a random labeled edge $s_1 = (e_1, e_2, R)$ satisfied by the patient's history that is connected to I and not in E. (To be connected to I, the new edge simply must contain a health event that is included in at least one of the edges in I.)

   (c) Create another labeled edge $s_2 = (e_1, e_2, R')$ where e_1 and e_2 are as in s_1 and R' is the temporal relation that e_1 simply precedes e_2.

   (d) Set S = S $\cup$ $\{s_1, s_2\}$

3. For each edge **s** in S, calculate the node population's conditional entropy $H(1_{Stroke}|\mathbf{1_s}) \equiv -\sum_{\mathbf{x}\in\{0,1\},\mathbf{y}\in\{0,1\}} \mathbf{P}(\mathbf{1_s}{=}\mathbf{x}, \mathbf{1_S}_{stroke}{=}\mathbf{y}) log(\mathbf{P}(\mathbf{1_S}_{troke}{=}\mathbf{y} \mid \mathbf{1_s}{=}\mathbf{x}))$ where P denotes the empirical probability measure for the node population and 1_{Stroke} is an indicator variable that is 1 if a patient had a stroke and 0 otherwise.

4. Let s* be the most informative edge in S, i.e. **s***= $\text{argmax}_{s\varepsilon S}$ $H(1_{Stroke}|\mathbf{1}_s)$ is the edge that maximizes conditional entropy. Create two new child nodes N_1 and N_2 with N_1 defined by (I_1=I $\cup$ **s***, E_1=E) and N_2 defined by (I_2=I, E_2 = E $\cup$ **s***).

Generating candidate labeled edges, which we do in Step 2 of Algorithm 1, posed some difficulties. In their speech recognition application [7], Amit and Murua used

domain knowledge to pre-specify a manageable set of possible edges and selected candidate edges to split each node uniformly at random from this set. Their temporal relations all took the form of non-overlapping time windows (e.g. event e_i occurs between T_1 and T_2 milliseconds before event e_j).

In our application, there are on the order of 100 million possible pairs of health events from which to choose the vertices for an edge. The probability of choosing useful pairs uniformly at random is negligible. We also lacked the prior information required to pre-specify appropriate temporal relations. The nature of our problem suggests that the temporal relations ought to be at least somewhat flexible for several reasons. First, there is a lot of presumably uninformative between-person variation in the rate at which similar biological processes unfold. Second, there is also presumably uninformative variation in the time from occurrence of a health event to when it appears as an insurance claim and enters our database. Therefore we do not expect extremely restrictive temporal relations (e.g. event e_i occurs exactly 10 days before event e_j) to lead to good classifiers. But it is unclear how permissive the temporal relations should be.

As a solution to these problems, we allowed the data to dictate candidate edge generation. We set temporal relations to be random relaxations of precise relations actually observed between health event pairs in the history of some patient in the node being split. The process is described in detail in Algorithm 1A. As an additional safeguard against overfitting, for each health event pair we also included a candidate edge with the further relaxed relation that event 1 simply precedes event 2 as described in step 2(c) of Algorithm 1.

Algorithm 1A (Selecting a random edge from a patient in a node defined by edge sets I and E, step 2b of Algorithm 1):
An edge will be of the form {event_1,event_2,(b,d)}, meaning event_1 occurs between 0 and b days before $\text{event}_2$and at most d days before index afib diagnosis.

1. Select a pair of health event occurrences e_1 and e_2 uniformly at random from the patient's health history such that at least one of the events is contained in some edge in I and the two event occurrences are not connected by any edge in E. Without loss of generality, let $\text{e}_1$be the earlier of the two events. Set e_1 and e_2 to be the vertices of the edge.

2. Let b' = $\text{t}_1 - \text{t}_2$ where t_i = the number of days before index that e_i occurred. Set $b \sim uniform(b', T)$ where T is the total number of days in the health history.

3. Let d' = t_1. Set $d \sim uniform(d', T)$

Algorithm 1A generates random edges by beginning with an 'exact' edge observed in a patient history and then relaxing the temporal restriction by a random amount. When growing a random forest, the goal is to find a good tradeoff between making each random tree a strong classifier and ensuring that the trees are as uncorrelated with each other as possible [5,6,15]. Our scheme for generating random edges is an intuitive attempt to find this tradeoff. Judgment and arbitrariness are still involved in selecting the general types of temporal relations we search for. It would be possible to consider many other relations than those we considered (for instance, the duration of time that two event eras overlap).

We also note that we allowed the selection of edges containing the index afib diagnosis as a vertex. Since every patient had an afib diagnosis on the last day of their history, such edges contain no relational information. By allowing them, we give RRF access to the sort of non-relational variables that could be included in a standard random forest.

To grow an individual tree, we split nodes until a stopping rule is satisfied. There are many reasonable options for stopping rules. For example, one can split nodes until the gain in entropy or the size of the node population falls below a pre-specified threshold. We continued to split nodes until their population fell below 100 patients.

Figure 3 depicts an example of a simple tree that might be grown using this algorithm. Figure 4 depicts an actual tree that was grown.

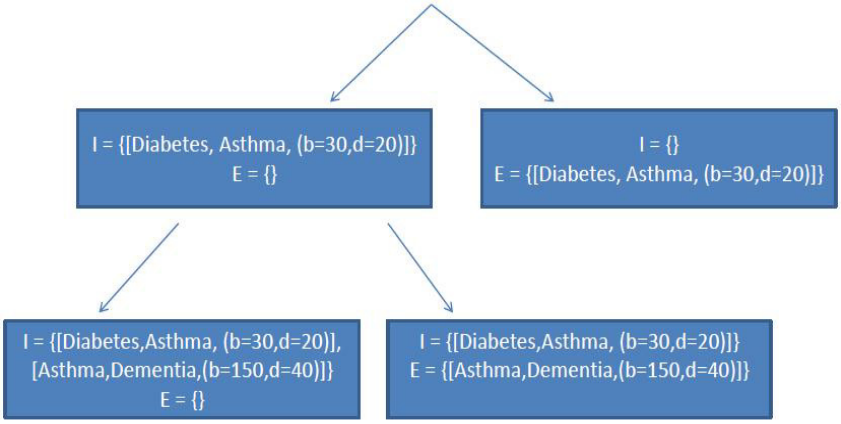

Figure 3.3: A hypothetical tree

As in [5,6,7], we grow many trees, splitting nodes as described in Algorithms 1 and 1A. The results we present are based on an ensemble of 1,000 trees. We determine the predictive score for an out of sample patient by dropping that patient down each tree in the ensemble and computing the pooled proportion of strokes in all the terminal nodes to which that patient belongs. Standard variable importance measures for conventional random forests [5] may be applied to identify the most predictive health event patterns.

We evaluate out of sample predictive performance of all methods using 4 fold cross validation. We consider summaries of predictive performance such as AUC scores and calibration plots and also examine the methods' ability to identify very low risk patients.

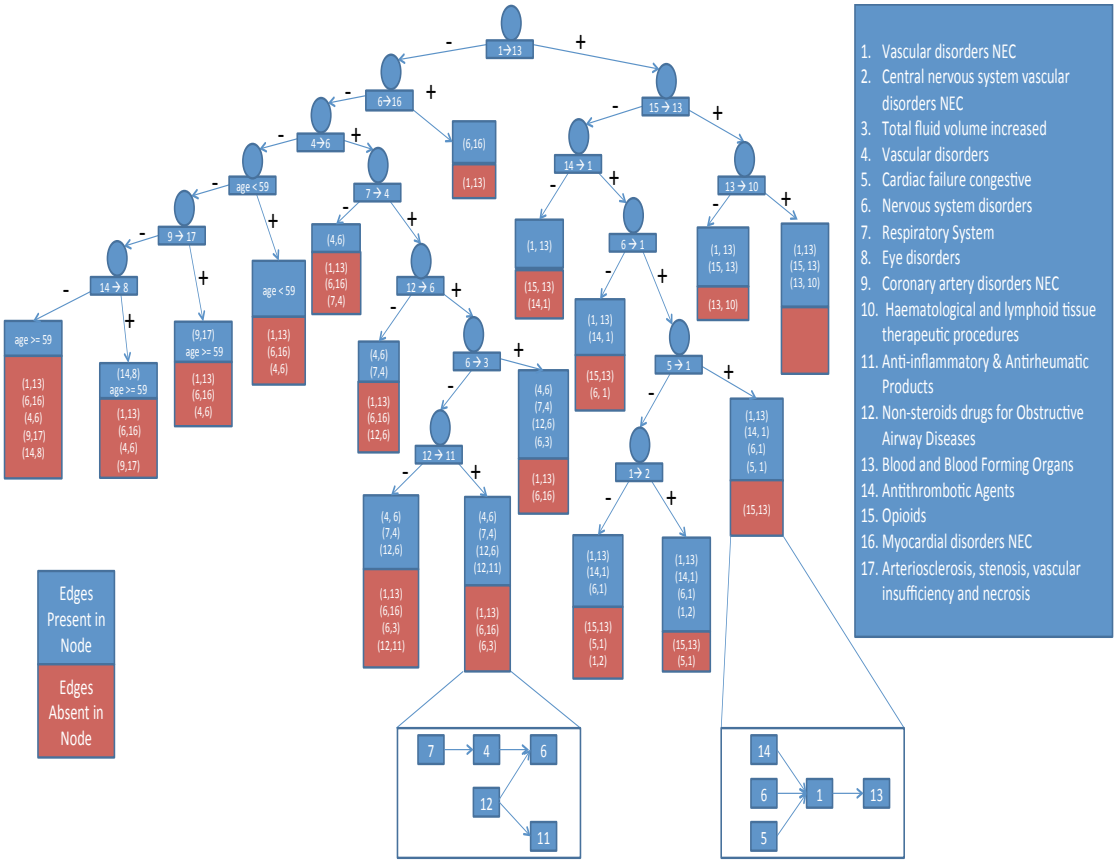

Figure 3.4: One actual tree from a Relational Random Forest. In this tree, for reasons of presentation, only ordering of events are depicted. At each terminal node we list the included and excluded edges. For two terminal nodes we illustrate the graph of temporal relations corresponding to the set of included edges. To the right is a key of health events that appear in the tree.

3.3 Results

The three machine learning methods had similar out of sample (OOS) AUCs and all outperformed CHADS2. (See Table 3.) They were also all well calibrated in that empirical probability of stroke was an approximately non-decreasing function of predictive score as illustrated in Figure 5. However, RRF was the only method

that was able to discriminate among low risk patients to identify a sizable population of very low risk patients (see second column of Table 3).

The calibration plots (Figure 5) illustrate the shared strengths of the three machine learning methods and the advantage of RRF. In each calibration plot, each point corresponds to a bin covering 2 percentiles of the model's standardized predictive score outputs. Each point's x-coordinate is the center of its bin, and the y-coordinate is the proportion of patients whose OOS predictive score falls within that bin who had strokes. Each calibration plot also includes a smoothed LOESS line fit to the binned proportions.

The calibration plots show that all three methods discriminate well among high risk patients. The nearly horizontal segment at the lower left of the random forest plot (top) indicates that the random forest does a poor job at discriminating among low risk patients, estimating that they all have approximately .05 probability of stroke. The model is still well calibrated but does not reliably discern gradations of risk below a certain threshold. The tightly clustered, football-shaped collection of points in the lower left corner of the logistic regression plot (middle) indicates that logistic regression does somewhat better at discriminating among low risk patients as it has a higher slope than the corresponding region of the random forest plot, but it still does not do well. In the corresponding area of the RRF plot (bottom), we see that the cluster from the logistic regression plot has been teased out and rises with a sharper slope as RRF does distinguish well between low and very low risk patients. Below, we look at some quantitative consequences of this visual comparison.

[10] suggested that patients with less than .02 probability of stroke in the next year in the absence of warfarin could reasonably be spared warfarin. In the studies [10] considered, this risk level corresponded to a CHADS2 score of 0. For each model we implemented, we looked at the proportion of the population (N = 12,581 patients) the model could identify with empirical probability of stroke less than p for various values of p including .02. To elaborate, for each model we found the highest predictive score $\tilde{p}$ such that less than $100 \times p\%$ of patients with OOS predictive score lower than $\tilde{p}$ had strokes. If there was no value of $\tilde{p}$ satisfying this condition, we set $\tilde{p} = -\infty$. We then considered the proportion of patients with OOS predictive score less than $\tilde{p}$ to be the proportion of patients with empirical risk of stroke less than p identified by the model. For example, 3.1% of patients with a CHADS2 score of 0 had strokes, and 11% of patients had CHADS2 scores of 0. So CHADS2 identified 11% of the population with empirical risk of stroke less than .031 but 0% of the population with empirical risk of stroke less than .02. As illustrated in Table 3, only RRF identified a sizable proportion of the population with empirical risk of stroke less than .02.

RRF also required far fewer trees to achieve its peak predictive performance than did the standard random forest in our application. Figure 6 shows AUC and proportion of low risk patients identified as functions of number of trees for both methods. The computational gain that may be accrued by using fewer relational trees may be offset by additional computational complexity involved in building each relational tree, however. Evaluating whether a patient's health history sat-

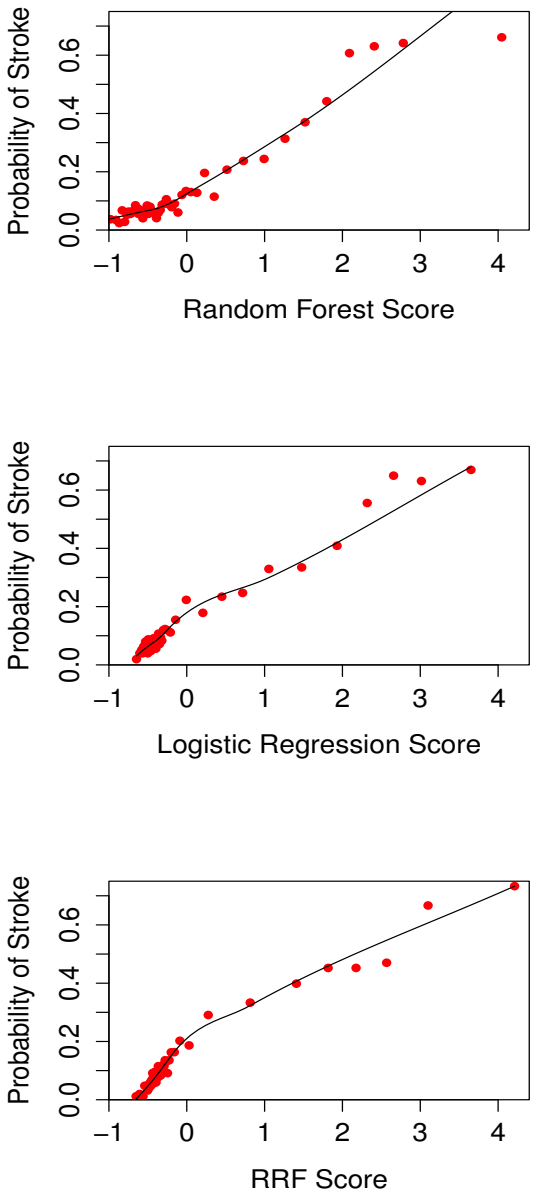

Figure 3.5: Calibration plots for the L1 Logistic Regression, Random Forest, and RRF classifiers.

Method	AUC	Proportion Identified With <2% Empirical Risk of Stroke	Proportion Identified With <1% Empirical Risk of Stroke
Chads2	.72	0	0
L1 Logistic Regression	.776	.005	.004
Random Forest	.782	0	0
RRF	.783	.1	.04

Table 3.2: Summary of predictive performance of various methods at predicting strokes. All machine learning methods had similar AUCs, but only RRF could identify a sizable proportion of the population at very low risk of stroke.

isfies an edge is a somewhat more complex operation than simply looking up the value of a pre-computed binary indicator variable. The worst case time complexity of splitting one node in RRF is $O((M+2) \times d \times N)$ where M is the number of candidate edges, d is the maximum number of occurrences of an event in any single patient's health history, and N is the number of patients in the node [7]. The time complexity for splitting a node in a random forest with binary predictors is $O((M+2) \times N)$. So the node splitting complexity of RRF is greater than standard random forests by a factor of d. In practice, d will often be quite small, as it was in our application where most events occurred very infrequently. In some applications, only the first occurrence of an event will be of interest and d will be 1. Of course, the theoretical upper limit on d is the number of time points at which each patient is observed.

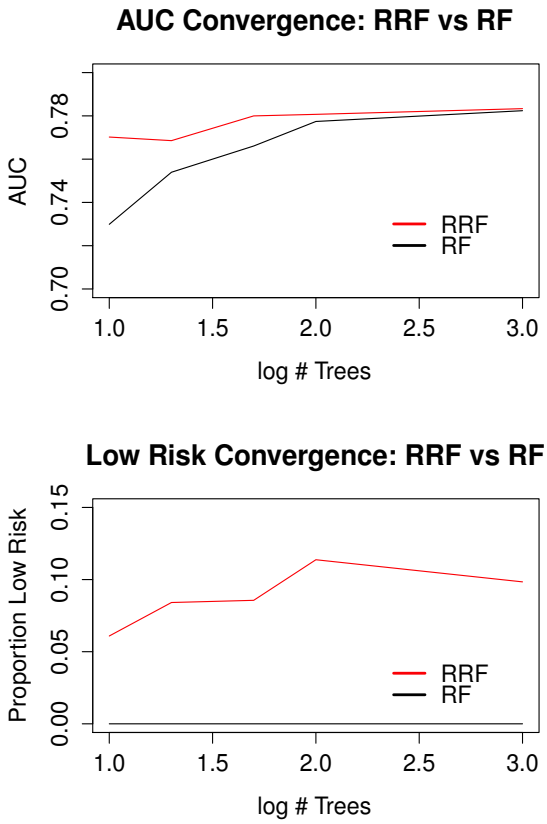

Figure 3.6: Predictive performance as a function of $log_{10}(\#$ trees) for RRF and standard random forest

3.4 Discussion

This application provides several heartening examples of predictive performance improving with utilization of additional information. The machine learning methods that use data related to all health events perform markedly better than CHADS2, which only takes a few researcher-selected health events as inputs. Of the machine learning methods considered, RRF was able to incorporate the most temporal information. The predictors that we constructed for logistic regression and random forest did contain coarse temporal information about the occurrence of individual health events, but only RRF could efficiently incorporate information from the vast space of temporal interactions among multiple health events into its predictions. For the task of predicting strokes from health histories considered here, this led to potentially meaningful improvements over the more conventional methods. More generally, RRF is a promising tool for predictive modeling in situations where the covariates comprise high dimensional time series.

A limitation of this study is the possibility that the probability estimates of all methods we considered were biased by the exclusion of patients who left Medicaid less than one year after their index afib diagnoses. Even if such bias was present, the classification problem we considered was still well defined, though of unclear clinical significance, and the predictive performance comparisons we presented are valid for that problem. However, clinical interpretation of the probability estimates would have to be relative to a somewhat unnatural population.

Though the emphasis of this paper is prediction, we briefly discuss some causal

considerations that naturally arise when using predictive models calibrated on observational data to aid clinical decisions. We adopt Pearl's do() operator [16] for explication. A full decision analysis of whether to prescribe warfarin to a patient with health history X would require the quantities P(stroke|X,do(no warfarin)), P(stroke|X,do(warfarin)), P(bleed|X,do(no warfarin)), and P(bleed|X,do(warfarin)). (Note that 'X' can vary depending on the predictors extracted from a health history.) CHADS2 aids clinical decision making by estimating just the first of these quantities – P(stroke|X,do(no warfarin)). CHADS2 is able to directly target this quantity because it was calibrated to patients randomly assigned not to take warfarin. Clinical considerations led experts to suggest that values of P(stroke|X,do(no warfarin)) below .02 imply potential gains from warfarin are so low as to not justify the risk. In the clinical trial population to which CHADS2 was calibrated, a CHADS2 score of 0 corresponded to a value of P(stroke|X,do(no warfarin)) that fell below this .02 threshold.

In our database of Medicaid patients, however, we saw that CHADS2 along with logistic regression and random forests were all unable to identify many patients with a marginal empirical probability of stroke P(stroke|X) $< .02$. The reasonable assumption that warfarin never increases the probability of stroke, i.e. that the quantity of interest P(stroke|X,do(no warfarin)) $>$ P(stroke|X) for all X, would imply that these methods are also unsuitable to identify patients from the Medicaid population with P(stroke|X,do(no warfarin)) $< .02$. Only RRF identified patients with a sufficiently low observational P(stroke|X) that they may possibly have prohibitively low causal P(stroke|X,do(no warfarin)) as well.

(We should note that CHADS2's relatively poor performance here may be due to the limited time horizon of our data. Patients only receive CHADS2 points for conditions related to insurance claims that occurred after they entered the database. This could result in some high risk patients wrongly being given scores of 0, thus inflating the proportion of strokes among those with scores of 0. Of course, the other methods we implemented were also hampered by left censoring of covariates.)

While we cannot reliably estimate the actual value of P(stroke|X,do(no warfarin)), we consider the conditions under which the quantile of observational predictive score P(stroke|X) at least corresponds to the quantile of P(stroke|X,do(no warfarin)). We start by asking when it can be that
P(stroke|X_1,do(no warfarin)) < P(stroke|X_2,do(no warfarin)) but
P(stroke|X_1) > P(stroke|X_2) for health histories X_1 and X_2. Assuming that warfarin always lowers the probability of stroke, this type of discordance can only occur if patients with history X_2 are more likely to take warfarin and/or warfarin is more efficacious for patients with history X_2. Whether patients with one predictive score are more likely to take warfarin than patients with another can be checked empirically. In our data, for example, we find no evidence that patients with very low RRF predictive scores were more likely to take warfarin than patients who had slightly higher predictive scores. Therefore, if the group of very low risk patients we identified does not correspond to the patients with the lowest P(stroke|X,do(no warfarin)), it is because warfarin was particularly efficacious in this group. If clinicians judge that such a pattern of warfarin treatment effect heterogeneity is implausible or unlikely then they can take the ordering of patients by

RRF predictive score as a proxy for the ordering of patients by P(stroke|X,do(no warfarin)).

To summarize, CHADS2's advantage of having been previously calibrated to a randomized trial is largely moot for the Medicaid population we consider given the failure of its calibration to randomized patients to generalize. Our predictive model does not directly estimate the desired quantity because it was fit to observational data, but it is still potentially a useful tool for identifying Medicaid patients who can be spared warfarin. Similar tools could be developed for other populations in which CHADS2 fails to identify low risk patients. Despite standard limitations of observational data, RRF represents a promising approach for making meaningful predictions from large-scale clinical databases.

3.5 References

[1] Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. 'Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership.' Ann Intern Med. 2010 Nov 2;153(9):600-6.
[2] Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. 'Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership.' Stat Med. 2012 Dec 30;31(30):4401-15.
[3] Ryan, PB et al. 'Studying the Science of Observational Research: Empirical Findings from the Observational Medical Outcomes Partnership.' Drug Safety, 2013, Vol. 36, Supplement 1 (pp. S1-S204).
[4] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc.B, 58, 267-288.
[5] Breiman, L. Random forests. Machine Learning, 45(1): 5-32, 2001.
[6] Amit, Y. and Geman, D., Shape quantization and recognition with randomized trees, Neural Computation, 9, 1545-1588, 1997

[7] Amit, Y. and Murua, A. 'Speech recognition using randomized relational decision trees,' IEEE Trans. Speech Audio Process. 9, 333-342, 2001
[8] He D., Mathews S.C., Kalloo A.N., et al. 'Mining high dimensional claims data to predict early hospital readmissions'. J Am Med Inform Assoc. Published Online First: 30 September, 2013
[9] Wang, F., Lee, N., Hu, J., Sun, J., and Ebadollahi, S. (2012, August). Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 453-461). ACM.
[10] Brian F. Gage, Carl van Walraven, Lesly Pearce, Robert G. Hart, Peter J. Koudstaal, B.S.P. Boode, Palle Petersen. 'Selecting Patients With Atrial Fibrillation for Anticoagulation Stroke Risk Stratification in Patients Taking Aspirin,' Circulation. 2004; 110: 2287-2292
[11] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. 'Validation of a common data model for active safety surveillance research'. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):54-60.
[12] BBR software retrieved from www.bayesianregression.org/bbr.html
[13] Genkin, A.; Lewis, D.; Madigan, D. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics, Vol. 49, No. 3, August 2007, pp.291-304
[14] Karampatziakis, N. (n.d.). FEST - Random Forests (and friends) for sparse data. Retrieved from Nikos Karampatziakis: http://www.cs.cornell.edu/ nk/fest/
[15] Murua, Alejandro. "Upper bounds for error rates of linear combinations of classifiers." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.5 (2002): 591-602.
[16] Pearl, Judea. Causality: models, reasoning and inference. Vol. 29. Cambridge: MIT press, 2000.

Chapter 4

A Note on the Effective Sample Size in Importance Sampling

4.1 Introduction

Importance sampling is a mainstay of scientific computing. Suppose we wish to evaluate the quantity:

$$\mu = \int_{\mathcal{X}} h(x)f(x)dx = E_f[h(X)],$$

where $\mathcal{X}$ is the support of the random variable X, $h(x) \geq 0$, and $f(\cdot)$ is a probability density function. Importance sampling draws $x^{(1)}, \ldots, x^{(m)}$ from an easy-to-sample trial distribution $g(\cdot)$ and estimates μ as:

$$\tilde{\mu} = \frac{1}{m}\left\{w^{(1)}h(x^{(1)}) + \ldots + w^{(m)}h(x^{(m)})\right\}$$

where $w^{(j)} = f(x^{(j)})/g(x^{(j)}), j = 1, \ldots, m$. Normalization of the weights is common and sidesteps the computation of normalizing constants in f and g. A good candidate for $g(\cdot)$ is one that is close to the shape of $f(x)$. However, so long as

the support of $g(\cdot)$ includes the support of $f(\cdot)$, essentially any distribution can be used. The literature includes an array of specialized importance sampling algorithms including sequential imputation (Kong *et al.*, 1994), adaptive importance sampling (Oh and Berger, 1992), and annealed importance sampling (Neal, 2001), to name a few.

While the variance of $\tilde{\mu}$ can in fact be smaller than that of an estimate obtained using independent samples from $f(\cdot)$ (Liu, 2001, p.24), it is more typically the case that importance sampling serves as a less precise substitute when $f(\cdot)$ is hard-to-sample. The variance of $\tilde{\mu}$ is sometimes large even for large m.

As a way of quantifying the difference in precision between estimates arising from different sampling methods, several authors have discussed rules of thumb to approximate the so-called "effective sample size" (ESS). The ESS of m importance sampling draws from $g(\cdot)$ to estimate μ is defined to be the number of independent draws directly from $f(\cdot)$ that would be required to obtain an estimate with the same variance. The standard formula for the approximate ESS is:

$$ESS(m) \approx \frac{m}{1 + \mathrm{Var}_g\left[w(x)\right]}. \tag{4.1}$$

The first explicit reference to the formula seems to be Kong *et al.* (1984), and it is apparently in widespread use.

Liu (1996) and especially Liu (2001) sketch a derivation of (1). Here we provide a detailed derivation and analysis of the approximation. Kong (1994) notes that the approximation is somewhat crude but still finds it appealing because it does not depend on $h(\cdot)$. Liu (2001) comments that one of its omitted terms "is not necessarily small." We show with some simple numerical examples that

Liu is indeed correct and that the formula can be misleading. We also propose a new approximation for the ESS and demonstrate its superior accuracy in some examples.

4.2 An Analysis of the Effective Sample Size Formula

4.2.1 Derivation

Importance sampling first generates independent samples $x^{(1)}, \ldots, x^{(m)}$ from an easy-to-sample trial distribution $g(\cdot)$ and then estimates $\mu = E_f[h(X)]$ by:

$$\hat{\mu} = \frac{\frac{1}{m}\sum_{j=1}^{m} h(x^{(j)})w(x^{(j)})}{\frac{1}{m}\sum_{j=1}^{m} w(x^{(j)})} = \frac{\tilde{\mu}}{\overline{W}}, \tag{4.2}$$

where $w(x^{(j)}) = f(x^{(j)})/g(x^{(j)}), j = 1, \ldots, m$.

Suppose we could also estimate μ with $\hat{\hat{\mu}} := \frac{1}{m'}\sum_{j=1}^{m'} h(y^{(j)})$ where the $y^{(j)}$ are i.i.d. draws from $f(\cdot)$. Then the ESS is, by definition, the value of m' for which $Var_f(\hat{\hat{\mu}}) = Var_g(\hat{\mu})$. Because $Var_f(\hat{\hat{\mu}}) = \frac{1}{m'}Var_f(h(y)), y \sim f$, we have:

$$ESS = m' = \frac{Var_f(h(y))}{Var_g(\hat{\mu})} \tag{4.3}$$

Consider first the denominator $Var_g(\hat{\mu})$ in (4.3). We will use Cramer's Theorem (Sen and Singer, 1993) to approximate this quantity:

Cramer's Theorem Let $\{T_m\}$ be a sequence of random q-vectors with common mean vector θ, r a positive integer, and $\pi\colon \mathbf{R}^q \to \mathbf{R}$ such that the following three conditions hold:

(i) $|\pi(T_m)| \leq cm^p$ where c and p are finite constants.

(ii) there exists $k > r(p+1)$ and $m_0 = m_0(k)$ such that for all $m \geq m_0$, $E[\sqrt{m}(T_{mj} - \theta_j)]^{2k} < \infty, j = 1, \ldots, q$

(iii) $\frac{\partial^2}{\partial x_i \partial x_j}\pi(\mathbf{x})$ is continuous and bounded in some neighborhood of θ,$i, j = 1, \ldots, q$

Then,

$$E\{\pi(\mathbf{T_m}) - \pi(\theta)\}^r = E\{\textstyle\sum_{j=1}^q \pi'(\theta)(T_{mj} - \theta_j)\}^r + O\left(m^{-\frac{r+1}{2}}\right).$$

Let $Z = h(x)w(x), W = w(x), H = h(x)$ and let $\overline{Z}$ and $\overline{W}$ be the corresponding sample averages, so that $\hat{\mu} = \frac{\overline{Z}}{\overline{W}}$. Define $\mathbf{T_m} := \left(\overline{Z}, \overline{W}\right)$. Then $\theta := (\mu, 1) = E_g(\mathbf{T_m})$. Define $\pi(a, b) = \frac{a}{b}$ so that $\pi(\mathbf{T_m}) = \frac{\overline{Z}}{\overline{W}} = \hat{\mu}$, $\pi(\theta) = \mu$, and $\pi'(\theta) = (1, -\mu)$. Plugging $\mathbf{T_m}$, θ, and π into Cramer's Theorem twice (once with $r = 2$ and once with $r = 1$) and assuming the theorem's regularity conditions hold yields:

$$E_g(\tfrac{\overline{Z}}{\overline{W}} - \mu)^2 = E_g[(\overline{Z} - \mu) - \mu(\overline{W} - 1)]^2 + O(m^{-\frac{3}{2}}) = E_g(\overline{Z} - \mu\overline{W})^2 + O(m^{-\frac{3}{2}})$$

and

$$E_g(\tfrac{\overline{Z}}{\overline{W}} - \mu) = E_g(\overline{Z} - \mu\overline{W}) + O(m^{-1}) = \mu - \mu + O(m^{-1}) = O(m^{-1}).$$

So:

$$Var_g(\hat{\mu}) = Var_g(\tfrac{\overline{Z}}{\overline{W}}) = Var_g(\tfrac{\overline{Z}}{\overline{W}} - \mu) = E_g(\tfrac{\overline{Z}}{\overline{W}} - \mu)^2 - [E_g(\tfrac{\overline{Z}}{\overline{W}} - \mu)]^2$$
$$= E_g(\overline{Z} - \mu\overline{W})^2 + O(m^{-\frac{3}{2}}) - O(m^{-2}) = Var_g(\overline{Z} - \mu\overline{W}) + O(m^{-\frac{3}{2}})$$

$$\approx Var_g(\overline{Z} - \mu\overline{W}) = \frac{1}{m}[Var_g(Z) + \mu^2 Var_g(W) - 2\mu Cov_g(Z, W)] \tag{4.4}$$

So far, we have ignored one remainder term,

$$R1 = Var_g(\frac{\overline{Z}}{\overline{W}}) - Var_g\{\overline{Z} - \mu\overline{W}\} = O(m^{-\frac{3}{2}}). \tag{4.5}$$

To reproduce Kong's formula, first note that

$$Cov_g(Z, W) = E_g(HW^2) - \mu = E_f(HW) - \mu = Cov_f(H, W) + \mu E_f(W) - \mu \tag{4.6}$$

Also,

$$\begin{aligned}
&Var_g(Z) = E_g(Z^2) - [E_g(Z)]^2 = E_g(H^2W^2) - \mu^2 = E_f(H^2W) - \mu^2 \\
&= E_f(H^2)E_f(W) + 2\mu E_f(HW) - 2\mu^2 E_f(W) + E_f(H^2W) - E_f(H^2)E_f(W) + \mu^2 E_f(W) \\
&\qquad -\mu^2 E_f(W) - 2\mu E_f(HW) + 2\mu^2 E_f(W) - \mu^2 \\
&= E_f(H^2)E_f(W) + 2\mu[E_f(HW) - \mu E_f(W)] + E_f\{[W - E_f(W)](H^2 - 2\mu H + \mu^2)\} - \mu^2 \\
&= [E_f(H)]^2 E_f(W) + Var_f(H)E_f(W) + 2\mu Cov_f(H, W) + E_f\{[W - E_f(W)](H - \mu)^2\} - \mu^2
\end{aligned} \tag{4.7}$$

Plugging (5) and (6) into (4), we get:

$$\begin{aligned}
&Var_g(\hat{\mu}) \approx \frac{1}{m}\{[E_f(H)]^2 E_f(W) + Var_f(H)E_f(W) + 2\mu Cov_f(H, W) \\
&+E_f\{[W - E_f(W)](H - \mu)^2\} - \mu^2 + \mu^2 Var_g(W) - 2\mu Cov_f(H, W) - 2\mu^2 E_f(W) + 2\mu^2\} \\
&= \frac{1}{m}\{\mu^2 E_g(W^2) + \mu^2 E_g(W^2) - \mu^2[E_g(W)]^2 - 2\mu^2 E_f(W) + Var_f(H)E_f(W) + \mu^2 \\
&\qquad +E_f\{[W - E_f(W)](H - \mu)^2\}\} \\
&\qquad = \frac{1}{m}Var_f(H)[Var_g(W) + 1] + \frac{1}{m}E_f\{[W - E_f(W)](H - \mu)^2\} \\
&\qquad\qquad \approx \frac{1}{m}Var_f(H)[Var_g(W) + 1].
\end{aligned} \tag{4.8}$$

We have now ignored a second remainder term,

$$R2 = \frac{1}{m} E_f\{[W - E_f(W)](H - \mu)^2\}, \tag{4.9}$$

to obtain $Var_g(\hat{\mu}) = \frac{1}{m} Var_f(H)[Var_g(W) + 1] + R1 + R2$.

Plugging this approximation into the expression for the true ESS (3), we get Kong's approximation:

$$ESS = m' = \frac{Var_f(H)}{Var_g(\hat{\mu})} \approx \frac{m}{Var_g(W) + 1}.$$

4.3 Numerical Study

To assess the performance of the approximation, we need to evaluate the true ESS ($\frac{Var_f(H)}{Var_g(\hat{\mu})}$). In the following examples, $f(\cdot)$ is simple enough to calculate $Var_f(H)$ exactly. We estimated $Var_g(\hat{\mu})$ by taking the sample variance of $\hat{\mu}'$s generated from many independent simulations.

4.3.1 Comparison of Formula with Truth for Some Simple Examples

Tables 4.1, 4.2, and 4.3 present simple univariate examples illustrating the (in)accuracy of Kong's formula. Table 4.1 considers two examples that simulate from an exponential distribution (specifically an exponential with parameter 1) when the target distribution is gamma. In one case (gamma(2,0)), the approximation overestimates the true ESS while in the other case (gamma(3,4)), the approximation underestimates the true ESS. In both cases the approximation does not appear to

	$f(x) \sim Gamma(a,b), g(x) \sim Exp(1)$			
	a=2 b=0.8		a=3 b=4	
m	True	Formula	True	Formula
100	8	26	127	68
500	42	132	627	342
1000	78	264	1224	684
2000	142	527	2616	1368
5000	311	1318	6405	3419
8000	516	2109	9420	5471
10000	627	2637	12403	6839
20000	1217	5373	24542	13678
50000	2739	13184	58974	34194

Table 4.1: Two examples comparing the true and approximate effective sample size for estimation of a gamma mean using an exponential as a trial distribution.

improve as the number of draws m increases. The two examples in Table 4.2 use a uniform distribution on zero to one as the trial distribution with a beta distribution as the target. In both cases, the approximation underestimates the true ESS. Finally, Table 4.3 shows two examples where both the trial distribution and the target distribution are normal. In one case the approximation underestimates the true ESS while in the other case it overestimates.

These tables clearly demonstrate that the approximate ESS can be quite far from the true ESS in either direction.

4.3.2 Remainders

We will show that disparities which persist even for large samples, e.g. in the tables above, are probably due to relatively large values of $R2$. We have that

$$ESS = \frac{Var_f(H)}{\frac{1}{m}Var_f(H)[Var_g(W)+1] + R1 + R2} = \frac{Var_f(H)}{\frac{1}{m}\left\{Var_f(H)[Var_g(W)+1] + mR1 + mR2\right\}}. \tag{4.10}$$

	$f(x) \sim Beta(a,b), g(x) \sim Unif(0,1)$			
	a=2 b=1		a=3 b=2	
m	True	Formula	True	Formula
100	95	75	108	73
500	494	375	544	365
1000	985	750	1095	729
2000	2000	1500	2205	1458
5000	4745	3750	5444	3646
8000	7208	6000	8775	5833
10000	9831	7500	10923	7292
20000	18951	15000	22037	14583
50000	41572	37500	54865	36458

Table 4.2: Two examples comparing the true and approximate effective sample size for estimation of a beta mean using a uniform as a trial distribution.

	$f(x) \sim N(a,b^2), g(x) \sim N(0,1)$			
	a=1 b=1		a=0.5 b=0.8	
m	True	Formula	True	Formula
100	22	37	95	78
500	93	184	463	388
1000	232	368	939	776
2000	369	736	1869	1553
5000	941	1839	4693	3881
8000	1482	2943	7389	6210
10000	1942	3679	9572	7763
20000	3930	7358	21037	15526
50000	9263	18394	49210	38815

Table 4.3: Two examples comparing the true and approximate effective sample size for estimation of a normal mean using a standard normal as a trial distribution.

Hence, for (1) to be a good approximation, R1 and R2 must both converge to 0 faster than $O(\frac{1}{m})$.

We saw in the derivation of the approximation in Section 2.1 that, if the conditions of Cramer's Theorem are satisfied, $R1 = O(\frac{1}{m^{\frac{3}{2}}})$. Condition (i) is easily satisfied if $h(\cdot)$ is bounded, which for any practical purpose it could be, even if by an extremely large number. Condition (ii) follows from Hoeffding's Inequality, which implies that the tail probabilities of $\sqrt{m}(\overline{Z}_m - \mu)$ and $\sqrt{m}(\overline{W}_m - 1)$ decay exponentially. For condition (iii), in our case $\pi''(\mathbf{x}) = (0, -\frac{1}{\overline{W}^2}, -\frac{1}{\overline{W}^2}, \frac{2\overline{Z}}{\overline{W}^3})$, which is continuous in a neighborhood of $\theta = (\mu, 1)$. Hence, as all the conditions of Cramer's Theorem are satisfied, $mR1$ should be small for large enough m. $mR2 = E_f[(W - E_f(W))(H - \mu^2)]$, on the other hand, is constant and not necessarily small compared to $Var_f(H)[Var_g(W)+1]$. In the example on the right half of Table 1, for instance, $Var_f(H)[Var_g(W) + 1] = 0.27$ and $mR2 = -0.24$.

4.4 An Adjusted Formula

As a remedy, we propose simply including $R2$ in the approximation, resulting in the formula

$$ESS \approx \frac{m}{Var_g(W) + 1 + \frac{mR2}{Var_f(H)}} \tag{4.11}$$

We can't evaluate $\frac{R2}{Var_f(H)}$ exactly in most cases, so we estimate it using importance sampling. This can be done without generating any additional samples beyond those already used for the primary estimation problem. We compared the performance of our adjusted approximation to the standard approximation in several examples for which we could determine the true ESS.

4.4.1 Numerical Study

Tables 4.4, 4,5, and 4.6 repeat the examples of Tables 4.1, 4.2, and 4.3, but now including the estimated ESS using the adjusted formula. In the tables, the "True ESS" column shows the actual ESS. "Kong's Formula" uses (1) to approximate the ESS. "Adjusted (exact)" uses the adjusted formula with an exact computation of R2. "Adjusted (simulated)" approximates R2 using the importance sampling draws. While we of course purposely selected examples where Kong's Formula fails, the examples that appear were the first and only considered for our modified estimator.

Note that even when importance sampling is inefficient, the adjusted approximation performs better than the standard formula.

m	$f(x) \sim Gamma(a = 3, b = 4), g(x) \sim Exp(1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	121	68	124	123	114	135
500	583	342	621	620	598	644
1000	1268	684	1241	1240	1208	1275
2000	2615	1368	2482	2481	2436	2529
5000	6271	3419	6206	6204	6132	6279
8000	10595	5471	9929	9928	9835	10025
10000	12938	6839	12411	12410	12307	12520
20000	24646	13678	24822	24825	24671	24970
50000	63842	34194	62056	62052	61823	62286

m	$f(x) \sim Gamma(a = 2, b = .8), g(x) \sim Exp(1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	8	26	6	21	7	43
500	42	132	28	67	22	123
1000	78	264	57	114	39	197
2000	142	527	113	199	72	330
5000	311	1318	283	421	173	670
8000	516	2109	452	653	227	986
10000	627	2637	565	806	353	1169
20000	1217	5373	1130	1500	662	2077
50000	2739	13184	2825	3409	1792	4548

Table 4.4: Two examples comparing the true and approximate effective sample size for estimation of a gamma mean using an exponential as a trial distribution.

m	$f(x) \sim Beta(a = 2, b = 1), g(x) \sim Unif(0, 1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	95	75	94	94	86	103
500	494	375	469	468	451	487
1000	985	750	938	937	911	962
2000	2000	1500	1875	1876	1839	1915
5000	4745	3750	4688	4686	4631	4742
8000	7208	6000	7500	7500	7428	7580
10000	9831	7500	9375	9378	9297	9454
20000	18951	15000	18750	18744	18626	18866
50000	41572	37500	46875	46872	46697	47065

m	$f(x) \sim Beta(a=3, b=2), g(x) \sim Unif(0,1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	111	73	109	109	102	117
500	553	365	547	547	532	564
1000	1095	729	1094	1093	1071	1118
2000	2106	1458	2188	2187	2156	2217
5000	5046	3646	5469	5469	5414	5518
8000	8974	5833	8750	8751	8685	8811
10000	10794	7292	10938	10939	10866	11007
20000	22656	14583	21875	21876	21777	21976
50000	55322	36458	54688	54680	54532	54838

Table 4.5: Two examples comparing the true and approximate effective sample size for estimation of a beta mean using a uniform as a trial distribution.

m	$f(x) \sim N(1,1), g(x) \sim N(0,1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	21	37	16	35	14	64
500	97	184	80	128	54	207
1000	198	368	160	240	111	358
2000	363	736	320	440	227	630
5000	924	1839	800	1032	609	1354
8000	1438	2943	1281	1607	1044	2020
10000	1755	3679	1601	1991	1377	2447
20000	3534	7358	3202	3881	2768	4664
50000	9277	18394	8005	9512	7536	10887

m	$f(x) \sim N(0.5, 0.8^2), g(x) \sim N(0,1)$					
	True ESS	Kong's Formula	Adjusted (exact)	Adjusted (sim median)	Adjusted (sim .05Q)	Adjusted (sim .95Q)
100	95	78	118	94	88	102
500	463	388	592	472	456	490
1000	939	776	1184	945	922	970
2000	1869	1553	2369	1889	1859	1924
5000	4693	3881	5922	4722	4670	4775
8000	7389	6210	9476	7555	7490	7621
10000	9572	7763	11845	9445	9373	9519
20000	21037	15526	23690	18888	18782	18990
50000	49210	38815	59225	47224	47069	47395

Table 4.6: Two examples comparing the true and approximate effective sample size for estimation of a normal mean using a standard normal as a trial distribution.

4.5 Conclusion

The ESS can be a useful tool for understanding the precision of estimates obtained through importance sampling. We have demonstrated that the standard approximation can be quite inaccurate, even for large numbers of draws, because it ignores a remainder term that is constant and not necessarily small. By simply estimating the remainder term, we can arrive at a better approximation. In situations where information conveyed by the ESS is consequential, it may be worth the additional computational cost to generate a more accurate approximation.

4.6 References

Kong, A., Liu, J.S., and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical*, **89**, 278–288.

Liu, J.S. (1996). Metropolized independent sampling with comparisons to re-

jection sampling and importance sampling. *Statistics and Computing*, **6**, 113-119

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer: New York.

Neal, R.M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**, 125–139.

Oh, M. S. and Berger, J. (1992). Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, **41**, 143–168.

Sen, P.K. and Singer, J.M. (1993) *Large Sample Methods in Statistics: An Introduction with Applications*. New York: Chapman and Hall